UNIVERSIDAD COMPLUTENSE DE MADRID

FACULTAD DE CIENCIAS ECONÓMICAS Y EMPRESARIALES
Departamento de Economía Cuantitativa

EVALUACIÓN Y COMPARACIÓN DE CAPACIDAD PREDICTIVA BAJO FUNCIONES DE PÉRDIDA DISCRETAS.

MEMORIA PARA OPTAR AL GRADO DE DOCTOR

PRESENTADA POR

Francisco Javier Eransus Armendáriz

Bajo la dirección del doctor

Alfonso Novales Cinca

Madrid, 2010

ISBN: 978-84-693-7810-6

TESIS DOCTORAL

EVALUACIÓN Y COMPARACIÓN DE CAPACIDAD PREDICTIVA BAJO FUNCIONES DE PÉRDIDA DISCRETAS

Francisco Javier Eransus Armendáriz

Director: Alfonso Novales Cinca

Universidad Complutense de Madrid
Facultad de Ciencias Económicas y Empresariales
Departamento de Economía Cuantitativa

Enero 2010

A mi padre,
de quien tanto aprendí,
a quien recuerdo cada día,
y *a quien sigo queriendo igual que cuando se fue*

AGRADECIMIENTOS

Este trabajo es el resultado de muchísimas horas de mi esfuerzo, constancia, aprendizaje y dedicación, pero no habría sido posible sin la ayuda y el apoyo de un grupo de personas que son especiales para mí y a quienes van dirigidas estas líneas, con todo merecimiento.

Antes de nada, debo mi agradecimiento sincero a mi director, Alfonso Novales. En él he hallado una fuente de conocimiento, de rigor científico y de honestidad intelectual, que han guiado mi tarea durante todos estos años. Pero más allá de su labor como director, Alfonso ha sido para mí un verdadero amigo, me ha protegido y me ha tratado con bondad, generosidad y cariño, sin esperar nada a cambio. Mi admiración por su grandeza profesional y humana es profunda.

En un plano similar debo colocar al profesor Carlos Sebastián, de quien he aprendido más economía en conversaciones de apenas minutos que en muchos cursos de estudio. Su labor durante años supervisando junto a Alfonso mi trabajo de previsión para la Comunidad de Madrid ha sido clave para que pudiera continuar en el mundo académico. Igual que ocurre con Alfonso, he tenido la suerte de que Carlos me permitiera disfrutar de su brillantez, de su claridad de ideas, de sus enseñanzas, y, sobre todo, de su amistad. Con él, he aprendido, me he reído y he sentido su afecto por mí, que es absolutamente recíproco.

Trabajar al lado de Alfonso Novales y Carlos Sebastián ha sido un auténtico honor.

Mi paso por el Departamento de Economía Cuantitativa de la Universidad Complutense me ha permitido conocer algunos magníficos compañeros, por cuyo apoyo en muchos facetas me siento profundamente agradecido. Lola Robles y Sonia Benito merecen un apartado destacado en este sentido. Este costoso trayecto ha sido bastante más ligero gracias a ellas, a su paciencia en escucharme, a sus consejos y su amistad. Pero, además, tengo bien presentes a Juan Angel, Rafi, Jesús, Sonia Brajín e Israel, Esther, Alfredo y Gustavo. A ellos debo muchos comentarios profesionales de interés y un gran número de inolvidables momentos personales.

Tampoco quiero dejar de resaltar la ayuda financiera y el comportamiento impecable que conmigo ha tenido la Consejería de Economía de la Comunidad de Madrid, en general, para quienes he trabajado y sigo trabajando con gusto, y, Conchy Ciruelos y Carlos Casado, en particular. Igualmente, agradezco las facilidades que el Departamento de Economía Cuantitativa de la Universidad Complutense de Madrid me ha otorgado para realizar esta investigación, y los medios que ha puesto a mi disposición para ello. Por último, me corresponde destacar explícitamente la concienzuda labor que la profesora Carmen García Olaverri ha realizado en la evaluación de este trabajo, y sus valiosos comentarios.

Si todos los mencionados hasta ahora han sido fundamentales para la realización de esta Tesis, punto y aparte merecen algunas personas que nada tienen que ver con el mundo de la universidad, pero mucho con mi vida. Primero, mis buenos y fieles amigos de Pamplona, en especial, Lion, Untxu y Eduardo. Sus visitas a Madrid forman parte de la historia de esta Tesis. Ellos me han aportado mucha alegría, muchas conversaciones de temas poco académicos y, más que nada, muchísimo cariño. Han asistido incrédulos y sorprendidos a mi marcha a la capital, y a mi consiguiente ausencia en infinidad de juergas y partidos de Osasuna, pero han aceptado, si bien no comprendido, que todo ha sido en nombre de la ciencia. Les debo mucho más de lo que ellos creen. No quisiera olvidar a otros buenos amigos pamploneses, como Villa, Iñigo e Ibon, Manolo y Rebeca, Sergio y Eva, Lasa y Chiara, César y Laura, Alfonso y Cristina, Luis, Cenzano y Arantxa, Txuma y Erviti, quienes, indirectamente, también han puesto su granito de arena. Además, Garbiñe merece una mención especial, por la preocupación que siempre ha mostrado por mis avatares universitarios, por los muchos ánimos que de ella he recibido en este tiempo y, en general, por su afecto.

En el sector madrileño, sin duda, a Ruth y a Valen corresponde mi gratitud sincera, por su comportamiento conmigo y por el interés y tolerancia que siempre han mostrado hacia mis extrañas conversaciones sobre economía y estadística. En tantas cenas en su casa, ellos han vivido de uno u otro modo todo el desarrollo de esta Tesis y han conocido de primera mano su evolución. Su apoyo moral en estos años ha sido muy relevante para mí, igual que lo es su leal amistad. Por otro lado, Angelines ha colaborado, con su cariño y con la bondad con la que siempre me ha tratado, a que me sintiera en Madrid como en mi propio hogar.

Como en las buenas novelas, dejo para el final lo más importante, mi agradecimiento a esas cinco personas especiales, sin las que ni este trabajo habría sido posible, ni en general mi vida tendría la misma ilusión ni el mismo color. Mi queridísima madre y mis hermanos Natxo y Jose, a los que adoro, me han animado, me han cuidado, me han escuchado y han sido un empuje fundamental para seguir adelante en los momentos difíciles. Pese a la inevitable distancia geográfica que en muchos tramos de cada año nos separaba, he sentido su cariño sin límites y su calor cada día. Ellos han dado por mí mucho más de lo que yo, por el tiempo absorto en este trabajo, he podido devolverles. Ojalá estas líneas os compensen mínimamente.

Si alguien ha vivido de lleno el largo proceso recorrido en esta investigación esa persona es Mariu. Ella ha sido mi sostén en todo momento, la fuente que me ha proporcionado durante estos años ánimo, paz, fuerza y alegría. Independientemente de cual fuera la situación, de las perspectivas que se atisabaran o de las contrariedades que surgieran, Mariu ha estado a mi lado, regalándome incondicionalmente su trato cálido, su sonrisa, su dulzura, su tiempo, su enorme corazón y su amor. Su presencia ha contribuido a este trabajo mucho más que todos los manuales de estadística y todos los teoremas aplicados, y, sobre todo, ha contribuido decisivamente a que, pese al esfuerzo invertido a veces con poca recompensa, me siguiera sintiendo feliz cada instante.

Y cierro esta larga y sentida lista de agradecimientos volviendo a mi padre, la persona a la que dedico esta Tesis como modesto pero merecidísimo homenaje a todo lo que él ha sido para mí. Pese a su ausencia, yo lo he sentido a mi lado durante todo este tiempo. Qué feliz hubiera sido viendo cómo este trabajo acaba con buen final, y cuánto hubiera disfrutado el día de su lectura. Ojalá puedas verlo desde donde estés, y te sientas tan orgulloso como yo me siento de tí.

Madrid, 24 de Noviembre de 2009

ÍNDICE GENERAL

INTRODUCCIÓN

Uno de los campos de mayor interés en casi cualquier disciplina científica es el relacionado con la previsión de los valores futuros que tomarán variables aleatorias relevantes ligadas a dicha materia, y esto es especialmente cierto en el caso de la economía. Una cuestión fundamental asociada con la previsión científica es el método o modelo para generar de forma adecuada dichas previsiones, que dependerá de la disciplina sobre la que se trabaja. Pero, una vez se generan las previsiones, otra cuestión importante es establecer una valoración sobre su calidad (e, indirectamente, sobre el método que las produjo), bien en términos absolutos, o bien comparativamente respecto a las previsiones alternativas ofrecidas por otra metodología. Para ello, en primer lugar, se requiere la definición de un criterio adecuado de valoración, que llamaremos "función de pérdida", que juzgue cada predicción puntual. Algunos ejemplos típicos en la literatura relacionada con la previsión económica son el cuadrado del error de previsión (SE) o su valor absoluto (AE),[1] y un listado de muchas otras funciones de pérdida menos convencionales pueden encontrarse, por ejemplo, en McCracken (2004). En segundo lugar, suponiendo que se dispone de uno o varios conjuntos de previsiones alternativos, formados cada uno por una serie temporal de predicciones sobre el valor que tomaría la variable en distintos momentos de tiempo, lo procedente es aplicar un test estadístico para contrastar formalmente la hipótesis que corresponda al respecto, en vez de hacer una mera comparación numérica de algún estadístico que resuma la secuencia de pérdidas obtenidas.

Este tipo de contrastes se dividen en dos categorías, según sea el objetivo perseguido: aquellos que evalúan si un conjunto de previsiones es "valioso" o "útil" para el usuario de las predicciones, y aquellos que comparan la precisión de dos (o más) conjuntos de previsiones que compiten entre sí. En el primer grupo se encuadran tests como el Test Binomial (B), el contraste de Henriksson-Merton (1981) (H-M), el de Pearson (Tabla de Contingencia (TC)) o el test de Pesaran-Timermmann (1992) (P-T). El problema inherente a esta tipología de contrastes es la ambigüedad de la hipótesis nula que subyace a ellos, obstáculo que se agrava, desde nuestro punto de vista, porque los procedimientos citados no introducen una función de pérdida. En el segundo grupo, la referencia fundamental es Diebold-Mariano (1995). Dicho test (DM) es válido asintóticamente bajo condiciones muy generales y para casi cualquier función de pérdida, pero la bondad de sus propiedades en muestras finitas fue probada en el artículo original solo bajo pérdida cuadrática (SE) y errores de previsión gaussianos.

Por otro lado, las previsiones realizadas podrían ser "ad-hoc" (previsiones basadas en juicios de valor e información cualitativa, sin usar modelos de predicción) o proceder de modelos cuyos parámetros toman valores numéricos conocidos. Éste es el contexto en el que fueron derivados originalmente los tests anteriormente citados y en el que se sitúa nuestro trabajo en los Capítulos 1 y 2. En concreto, el primero de estos capítulos trata sobre contrastes para la evaluación de un conjunto de previsiones, y el segundo, sobre tests para la comparación de la capacidad predictiva de dos conjuntos alternativos.

Sin embargo, normalmente, las previsiones proceden de un modelo estimado. En esa situación ("incertidumbre paramétrica"), las derivaciones originales de los contrastes anteriores no son correctas, ya que ignoran una parte de la varianza del estadístico de contraste debida a la estimación paramétrica. Existe una amplia literatura a este respecto, en la que destacan West (1996) y McCracken (2000), como referencias básicas. Pues bien, éste es el marco en el que se desarrolla el Capítulo 3, que pretende comprobar el efecto de la incertidumbre paramétrica en las propiedades del tipo de tests presentados en los dos capítulos precedentes, y realizar las correcciones oportunas en la varianza de los estadísticos de contraste, si procede, para que continúen siendo válidos.

Como ya se ha dicho, la evaluación de la precisión o calidad de las previsiones requiere la definición de un criterio o función de pérdida. En muchos contextos predictivos, las funciones convencionales (por ejemplo, SE) son inapropiadas para tal evaluación, pese a su uso casi sistemático en la literatura. Por ejemplo, si la variable a predecir es la tasa de crecimiento de un agregado macroeconómico, parece obvio que dos previsiones cuya distancia al verdadero valor del dato sea la misma, pero siendo una de ellas certera en signo y la otra no, no deberían valorarse por igual. En este trabajo, nosotros proponemos una clase de funciones de pérdida, que llamaremos "discreta". Dicha función consiste en particionar el dominio $R \times R$ de datos y previsiones en un número finito de cuadrantes y asignar una pérdida a cada uno de ellos.

[1] Las notaciones SE y AE se refieren a "Square Error" y "Absolute Error".

Las **ventajas** de esta clase de funciones son notables:

a) Desde un **punto de vista conceptual**:

a.1) No es una función de pérdida particular, sino que abarca toda una clase de ellas.[2] De este modo, es el usuario quien, según sea el ámbito predictivo concreto en el que se encuentre, elige la función adecuada, definiendo él mismo la partición y las pérdidas numéricas.

a.2) Subsanan las deficiencias asociadas a las funciones de pérdida convencionales, introduciendo un alto grado de flexibilidad en la valoración. En primer lugar, la función no se establece directamente sobre el error de previsión, sino, de forma más genérica, sobre el par dato-previsión. Esto permite, por ejemplo, evaluar también el signo de la previsión, y no solo la magnitud de la distancia entre ésta y el dato. En segundo lugar, admite la introducción de asimetrías en la valoración, como, por ejemplo, que un error de la misma magnitud se considere más grave cuando se produce en datos positivos que en datos negativos. Estas propiedades sugieren que este tipo de funciones pueden ser muy adecuadas para aplicarse con previsiones de variables económicas, especialmente agregados macroeconómicos.

a.3) Es idónea para la valoración de previsiones sobre variables de naturaleza cualitativa, sean o no de carácter económico.

b) Desde un **punto de vista técnico**, durante el desarrollo de este trabajo se muestra que, si se definen bajo funciones de pérdida discretas, los contrastes de evaluación o comparación de capacidad predictiva verifican las siguientes propiedades:

b.1) El efecto de valores atípicos sobre las propiedades de los tests en muestras finitas es nulo, algo que no se verifica bajo funciones de pérdida continuas.

b.2) El efecto de la incertidumbre paramétrica sobre las propiedades asintóticas de los contrastes tiende a ser, en general, mucho menor que bajo las funciones de pérdida estándar, lo que puede permitir en muchos contextos predictivos no tener que corregir la definición del contraste (ni la varianza del estadístico ni su distribución de probabilidad) respecto al caso de no incertidumbre paramétrica. Éste es, probablemente, *el resultado más relevante de la Tesis*.

Con todo lo dicho hasta este punto, ya estamos en condiciones de especificar el marco global en el que se encuadra esta Tesis, y sus objetivos generales. Nos moveremos en el ámbito de los contrastes estadísticos para la evaluación de la utilidad o valía de un conjunto de previsiones (Capítulo 1) y para la comparación de habilidad predictiva entre dos conjuntos alternativos (Capítulo 2), bajo funciones de pérdida discretas. En el primer caso (evaluación de un único conjunto de previsiones), propondremos una serie de tests sencillos que surgen de forma natural cuando el criterio de valoración de las previsiones es la función de pérdida discreta, y estudiaremos su comportamiento en comparación con los contrastes habituales de la literatura (B, H-M, TC, P-T), que no incluyen función de pérdida en su definición. En el segundo caso (comparación entre dos conjuntos de previsiones), trataremos de comprobar el tamaño y potencia del test DM en muestras cortas cuando la función de pérdida elegida es la discreta, y, además, sugerimos otro contraste que se establece exclusivamente bajo dicha función. En los dos primeros capítulos, supondremos que no existe estimación paramétrica en los métodos generadores de las previsiones. El objetivo del tercero será precisamente analizar las implicaciones de dicha estimación sobre las propiedades estadísticas del tipo de contrastes tratados en los capítulos anteriores.

El desarrollo de la Tesis se realiza de la siguiente manera:

El **CAPÍTULO 1** se ocupa del análisis de los contrastes de evaluación de la capacidad predictiva de un único conjunto de previsiones, en un entorno de no estimación paramétrica. Primeramente, se revisan los tests estadísticos utilizados en la literatura para este fin (B, H-M, TC y P-T), que se definen directamente sobre datos y previsiones, sin una función de pérdida explícita. Salvo el test B, el resto contrastan la hipótesis nula de "no utilidad de las previsiones" a través del enunciado formal "los datos y las previsiones son estocásticamente independientes". Argumentamos la inconveniencia de dicho enunciado y mostramos los efectos adversos que puede tener en las decisiones de rechazo o no rechazo de la hipótesis nula asociadas a los tests. Esto sirve para motivar la introducción de una función de pérdida discreta, que permite definir de forma natural contrastes alternativos, que no incurran en los errores que se describirán para los tests convencionales. El **objetivo del capítulo** es sugerir contrastes para evaluar la capacidad predictiva de un conjunto de previsiones, que se especifiquen bajo funciones de pérdida discretas, y estudiar

[2] No obstante, a lo largo de todo el texto de la Tesis, nos tomaremos la licencia de hablar de "función de pérdida discreta", y no de "clase de funciones de pérdida discretas".

su comportamiento. Se proponen tres contrastes: C1, en dos versiones (C1-v1 y C1-v2), C2 y C3. Se analizan sus resultados, comparativamente con los de los tests B, H-M, TC y P-T, a través de ejercicios de simulación. La peculiaridad de la hipótesis nula de los tests en este contexto no permite un análisis de tamaño y potencia convencional. Se utiliza un enfoque menos formal, que muestra los beneficios de aplicar tests como C1 ó C2. Al contrario que C2, tanto C1-v1 como C1-v2 pueden calcularse de forma sencilla y rápida.

El **CAPÍTULO 2** trata sobre contrastes para comparar la habilidad predictiva de dos conjuntos de previsiones alternativos, aplicados bajo funciones de pérdida discretas, de nuevo en el marco de inexistencia de estimación paramétrica. El **objetivo del capítulo** es doble:

1) Presentar estimaciones sobre las propiedades estadísticas en muestras cortas del test de referencia en este contexto, el test DM, cuando se aplica con una función de pérdida discreta. Téngase en cuenta que la distribución asintótica del test es conocida para una función de pérdida genérica, pero, por el contrario, sus resultados en muestras finitas pueden depender del tamaño muestral, de la distribución de los errores de previsión y de la función de pérdida.

2) Sugerir contrastes alternativos a DM para este contexto, y chequear si pueden ser ventajosos respecto a éste en algún caso.

Antes de pasar a resolver las cuestiones 1) y 2), se presentan estimaciones de tamaño y potencia para el test DM bajo pérdida cuadrática (SE) y bajo una función discreta, respectivamente, para mostrar evidencia sobre la ventaja b.1) asociada a las pérdidas discretas. Posteriormente, se sugieren tres contrastes alternativos a DM, válidos solo bajo pérdida discreta: Mult2, Mult2-aprx y RV-p, donde Mult2-aprx solo es la versión asintótica de Mult2. Igual que ocurre para los presentados en el Capítulo 1, estos contrastes se derivan de forma sencilla a partir del hecho de que las pérdidas discretas asociadas a cada conjunto de previsiones tienen un soporte finito y, por tanto, las frecuencias asociadas constituyen un vector aleatorio cuya distribución pertenece a la familia Multinomial. Se llevan a cabo ejercicios de simulación convencionales con los tres tests propuestos, y otros tres clásicos en la literatura, el test DM y los tests de Signos y Wilcoxon, aplicados también bajo función de pérdida discreta. Los resultados obtenidos muestran que las diferencias en el tamaño y potencia de los contrastes DM, Mult2 y Mult2-aprx son pequeñas en la mayor parte de diseños predictivos utilizados, si bien Mult2 es el más preciso en tamaño prácticamente en todos los casos. Como contrapartida, dicho test presenta un alto coste computacional. Se recomienda utilizar el test Mult2 en situaciones de autocorrelación en las pérdidas (lo que ocurre, por ejemplo, cuando el horizonte de predicción es mayor que uno) y longitud muestral pequeña, contexto en el que se identifican diferencias notables entre el sesgo en el tamaño de nuestro contraste y el del test DM, favorables al primero. Finalmente, todos los contrastes implementados bajo función de pérdida discreta presentan sesgos inaceptables en tamaño si las previsiones tienen ciertas características y la muestra es muy corta. El último punto del capítulo consiste en caracterizar dichas situaciones.

El **CAPÍTULO 3** contiene, probablemente, los resultados más relevantes de la Tesis. Como ya se ha anunciado anteriormente, esta parte del trabajo se enmarca en la literatura sobre contrastes de evaluación/comparación de capacidad predictiva bajo incertidumbre paramétrica (IP). Nuestro análisis se centra, por supuesto, en el efecto de la IP sobre dichos contrastes cuando éstos se implementan con una función de pérdida discreta. Las **preguntas fundamentales que se pretenden responder en el capítulo** son:

(i) ¿Cómo modificar los tests de los capítulos anteriores cuando nos encontramos en un contexto donde existe IP?.

(ii) En tal contexto, ¿existe alguna ventaja derivada del uso de una función de pérdida discreta para implementar dichos contrastes?.

Si las previsiones son generadas por un modelo cuyos parámetros se han estimado, como es habitual en la práctica, la distribución asintótica de los contrastes para la evaluación/comparación de capacidad predictiva no coincide con la presentada en los capítulos anteriores, porque aparece una *varianza adicional en el estadístico de contraste* debida a la estimación paramétrica del modelo, *que debe ser contabilizada.* La expresión analítica de dicha varianza depende del contexto predictivo y de la función de pérdida, fundamentalmente. Nos centraremos en el caso de que las variables que intervienen en los modelos de previsión sean estacionarias, y separaremos nuestro análisis según los modelos que compitan sean anidados o no anidados, ya que la teoría estadística al respecto difiere según esta característica.

En el marco de modelos **no anidados** (con variables estacionarias), las referencias fundamentales son West (1996) y McCracken (2000), que derivan la distribución asintótica bajo IP de una amplia familia de tests estadísticos implementados con función de pérdida diferenciable y no diferenciable en los parámetros del modelo, respectivamente. Entre ellos se encuentran los tests B, P-T, C1, DM y Mult2-aprx. La distribución derivada por West (1996) y McCracken (2000) resulta ser Normal, con una matriz de covarianzas que denotaremos por Ω, y que puede diferir notablemente de la matriz de covarianzas que se produce cuando no hay estimación paramétrica, que denotaremos por S_{ff} (ésta es la empleada en los Capítulos 1 y 2). La matriz Ω es la suma de S_{ff} y una composición de otras matrices, que dependen del modelo de previsión, de los métodos de estimación y de la función de pérdida utilizada en el contraste. Tal y como muestra West (1996), ignorar la corrección de la varianza asintótica, usando S_{ff} en vez de Ω, puede tener efectos dramáticos en las propiedades del test.

El problema de aplicar la varianza correcta Ω en los contrastes es su *dificultad de cálculo*. La estimación de Ω es tediosa en todos los casos, pero en muchas situaciones incluso puede no ser factible, lo que sucede si el usuario del test no conoce los detalles de la estimación paramétrica del modelo. Aún más, si la función de pérdida es no diferenciable, la obtención solamente de la expresión analítica de Ω puede ser una tarea complicada, en general. Precisamente, éste es el caso de nuestra función discreta, que no solo no es diferenciable, sino ni siquiera continua. En general, todos estos obstáculos son difícilmente esquivables. Sin embargo, en algunos ámbitos predictivos se verifica que $\Omega = S_{ff}$, es decir, que el efecto de la IP es nulo (irrelevancia asintótica de la incertidumbre paramétrica (IAIP)), lo que permite que el usuario del contraste se despreocupe de la molesta corrección de la varianza del test debida a la IP. Desgraciadamente, son pocos los casos en los que se ha identificado tal propiedad. Ésta aplica, por ejemplo, si la función de pérdida es SE ó AE y, además, se cumplen una serie de condiciones sobre el modelo de previsión y el método de estimación. Nuestro **objetivo** es triple:

1. Obtener una expresión analítica de Ω cuando la función de pérdida es discreta. Nos limitamos al caso de modelos lineales.

2. Si existe, caracterizar el conjunto de condiciones bajo las que la aplicación de los contrastes con una función discreta conduce a una situación de IAIP.

3. Obtener evidencia empírica (mediante simulación) respecto a la magnitud del efecto de la IP sobre los contrastes si se emplea (erróneamente) S_{ff}, en situaciones en las que la IP es relevante.

Las tres cuestiones son resueltas satisfactoriamente en la primera parte del capítulo. En primer lugar, se obtiene la expresión analítica general de Ω bajo la función de pérdida discreta, con la que, teóricamente, los contrastes B, P-T, C1, DM (bajo función discreta) y Mult2-aprx pueden extenderse al contexto de IP, siempre que se esté en un marco de modelos de previsión lineales no anidados y con variables estacionarias. En segundo lugar, se demuestra analíticamente que, bajo cierto conjunto de condiciones, prácticamente igual al que se requiere para tener la misma propiedad con función de pérdida AE, la aplicación de los contrastes bajo función de pérdida discreta garantiza IAIP. Este resultado está enunciado en el Corolario 1 del Capítulo 3, y constituye una de las conclusiones más relevantes de la Tesis. Finalmente, los ejercicios de simulación correspondientes al tercer objetivo sugieren que, en caso de que se incumplan las hipótesis que garantizan IAIP en modelos de previsión lineales no anidados con variables estacionarias, el sesgo en tamaño del test DM y de Mult2-aprx cuando no se corrige por la IP es muy inferior si el contraste se implementa bajo función de pérdida discreta que si se utiliza SE o AE, tanto en términos asintóticos como en muestras finitas, hasta el punto de que proceder de tal manera parece perfectamente razonable en la práctica, asumiendo un coste poco significativo en el comportamiento del test.

La segunda parte del capítulo estudia el efecto de la IP en los contrastes de comparación de capacidad predictiva cuando los modelos de previsión son **anidados** (también se restringe al caso de variables estacionarias).[3] En este contexto, no existe un resultado general sobre la distribución asintótica de esta clase de tests, análogos a los obtenidos por West (1996) o McCracken (2000) para modelos no anidados. En la literatura relativa a este asunto, se ha logrado derivar la distribución asintótica solo de algunos contastes en particular, y, además, bajo algún conjunto de supuestos. Las distribuciones resultantes suelen ser no estándar. La referencia más relevante es, probablemente, McCracken (2007), quien obtiene la distribución asintótica del test DM siempre que se aplique con una función de pérdida diferenciable y en una situación que verifique ciertas propiedades adicionales. Además de ser no estándar, la distribución depende del valor

[3] Al exigirse anidamiento, es obvio que se requieren (al menos) dos modelos de previsión. Por ello, los resultados se refieren a contrastes de comparación de capacidad predictiva, no de evaluación de ésta.

de dos parámetros, uno de los cuales mide en cierto modo el grado de IP existente. Ambas cuestiones incomodan la implementación del contraste. Nuestro **objetivo** vuelve a consistir en obtener evidencia sobre cuál es el efecto de la IP en este contexto si los tests de comparación de capacidad predictiva se aplican bajo función de pérdida discreta. No pretendemos lograr un enunciado teórico, sino que trataremos de inferir las conclusiones a través de ejercicios de simulación.

Los resultados obtenidos vuelven a ser muy positivos, ya que sugieren que, si el contraste DM se implementa con función de pérdida discreta, las propiedades del test se alteran escasamente en la mayoría de escenarios predictivos diseñados pese a la presencia de IP, sin necesidad de correcciones ni en la varianza del estadístico de contraste ni en la distribución de contraste empleada, al contrario de lo que ocurre con funciones de pérdida como SE. El resultado se verifica también para Mult2-aprx.

La conclusión global del Capítulo 3 es que el uso de una función discreta para valorar las previsiones proporciona un *elevado grado de robustez* en los contrastes de evaluación o comparación de capacidad predictiva *a la existencia de IP*. Es decir, en relación a las cuestiones (i) e (ii) planteadas anteriormente, nuestros resultados sugieren que los tests propuestos en los Capítulos 1 y 2 pueden aplicarse siguiendo su definición original, aun cuando las previsiones sean generadas por un modelo (lineal y con variables estacionarias) cuyos valores paramétricos se han estimado, sin necesidad de realizar correcciones (ni en la varianza del estadístico de contraste ni respecto a la familia de la distribución de contraste). Esta conclusión constituye la ventaja b.2) mencionada antes, respecto al empleo de funciones de pérdida discretas en contrastes de evaluación y comparación de precisión predictiva.

Aunque en cada capítulo se detallan las conclusiones fundamentales obtenidas en el mismo, todas éstas se recopilan de forma unificada en el último capítulo de la Tesis.

CAPÍTULO 1

CONTRASTES ESTADÍSTICOS PARA EVALUAR LA CAPACIDAD PREDICTIVA DE UN CONJUNTO DE PREVISIONES, BAJO FUNCIÓN DE PÉRDIDA DISCRETA

1. Introducción

Existen un buen número de procedimientos para evaluar la capacidad predictiva asociada a un conjunto de previsiones. La mayoría de las evaluaciones estadísticas se fundamentan en criterios de naturaleza cuantitativa, como el MSE, MAE o U-Theil, y existe una amplia discusión en la literatura sobre cuál de dichas medidas es mejor (por ejemplo, Amstrong y Fildes (1995), Baillie et al (1993), Clements y Hendry (1993)). Sin embargo, el uso de este tipo de criterios estándar para evaluar previsiones sobre variables económicas conlleva dos tipos de problemas. En primer lugar, los valores atípicos distorsionan de forma significativa el valor de estos estadísticos cuando la muestra manejada es corta, tal y como ocurre habitualmente. Y en segundo lugar, las medidas mencionadas son simétricas alrededor de cero, de modo que no tienen en cuenta el signo del error sino solo su cuantía. Esto último parece poco razonable para una buena parte de las variables económicas sobre las que se hace previsión. Las variaciones en tipos de interés, las variaciones en el precio de la mayoría de activos financieros y los cambios entre dos tasas consecutivas de crecimiento interanual en series macroeconómicas (especialmente agregados de contabilidad nacional) y en índices de precios de consumo, constituyen ejemplos obvios de la relevancia del signo de la previsión.

Así, en los últimos 20 años ha surgido una nueva línea de investigación y análisis empírico en la literatura sobre evaluación de previsiones, centrada en el uso de criterios estadísticos cuyas valoraciones son de tipo cualitativo, en cuanto a capacidad de predecir correctamente el signo o dirección del cambio en la variable, mientras poco o ningún peso es otorgado a la magnitud o tamaño del error cometido. Un gran número de trabajos en esta línea se han producido en los últimos años, tanto a nivel metodológico como aplicado. Ejemplos de los primeros son Merton (1981), Henriksson y Merton (1981), Schnader y Stekler (1990) y Pesaran y Timmermann (1992, 1994), especialmente, mientras de entre los segundos cabe destacar los artículos de Cumby y Modest (1987), Schnader y Stekler (1990), Stekler (1994), Leitch y Tanner (1995), Kolb y Stekler (1996), Ash, Smyth y Heravi (1998), Mills y Pepper (1999), Joutz y Stekler (2000), Oller y Bharat (2000), Pons (2000) y Greer (2003).

Este tipo de análisis direccional (terminología utilizada, por ejemplo, en Ash, Smyth, Heravi (1998)) es teóricamente aplicable a una amplia gama de datos susceptibles de predicción estadística, pero es especialmente adecuado para determinadas series de carácter económico, bien series financieras o bien macromagnitudes, como hemos mencionado. Respecto a las primeras, porque las estrategias de mercado, para ser exitosas, esencialmente requieren que las previsiones sean certeras solamente respecto a la dirección en la que se moverán las variables en cuestión. Leitch y Tanner (1991) aportan evidencia empírica en el sentido de que existe muy débil correlación entre beneficios económicos obtenidos de la aplicación de determinadas estrategias en el mercado de deuda pública que requerían una previsión y los valores de estadísticos de medición del tamaño de error (emplean, en concreto, MAE, RMSE y U-Theil) y, por el contrario, alta correlación entre dichas ganancias y los valores de criterios que medían la precisión respecto a la previsión del cambio de dirección de los tipos de interés. Por lo que respecta a series macroeconómicas, la dificultad de predecir con exactitud los valores de las tasas de crecimiento de series como los agregados de contabilidad trimestral incluso en el corto plazo y, sobre todo, el alto interés que tiene anticipar si la variable o la economía se está acelerando o desacelerando, justifican el uso de métodos cualitativos para valorar las previsiones respecto de tales tipos de datos. Birchenall et al (1996) subraya que "puede ser muy difícil hacer afirmaciones de tipo cualitativo respecto del estado de la economía, incluso cuando estas afirmaciones son del tipo 'la economía se está expandiendo' o 'la economía se está contrayendo'", e incide en que "no obstante, tal información cualitativa es importante prerrequisito para la correcta implementación de política monetaria y fiscal". Según Leitch y Tanner (1995), "si la economía va a crecer más rápidamente en el próximo trimestre o año que lo hizo en el actual o anteriores, ¿es relevante conocer si la tasa será exactamente de 5 % ó 10 %?", "la primera y fundamental cuestión es si la economía se está acelerando o desacelerando y la segunda, y probablemente, secundaria, cuánto se acelera o desacelera". Cualquier análisis de coyuntura actual está en sintonía con los comentarios anteriores.

Los trabajos de Cumby y Modest (1987), Kolb y Stekler (1996) y Greer (2003) son llevados a cabo sobre variables financieras, mientras los de Schnader y Stekler (1990), Stekler (1994) y Ash, Smyth y Heravi (1998), por ejemplo, aplican el análisis direccional sobre series macroeconómicas (en el artículo de Ash, Smyth y Heravi (1998) el estudio también se extiende a índices de precios de consumo, en concreto a la primera diferencia de la tasa de inflación).

Los tests de hipótesis han constituído el modo fundamental en que se han evaluado previsiones respecto a cambio direccional, contrastando hipótesis sobre utilidad o no de las previsiones (en cierto sentido que explicamos después). Hasta la fecha, son esencialmente cinco los tests propuestos y utilizados en la literatura. Merton (1981) y Merton y Henriksson (1981) idearon el primer contraste (H-M) formal en este contexto, y Schnader y Stekler (1990) presentaron un test de Pearson (TC 2×2) equivalente asintóticamente al anterior, en versión Tabla de Contingencia 2×2, mientras Greer (2003) sugiere y aplica un sencillo test basado en la distribución Binomial (B).[1] Estos tres contrastes emplean una partición del dominio de los datos y previsiones en dos regiones, separadas por un valor numérico razonable según el caso, que, habitualmente, será el cero (Schnader y Stekler (1990) presentaron también una aplicación donde el valor que divide en dos el espacio de posibles datos, GNP en su caso, era 2 %), de modo que los citados tests se fundamentan únicamente en el acierto de las previsiones respecto del signo. Sin embargo, es obvio que en muchas situaciones la información que se pretende predecir podría ser susceptible de separarse en más de dos categorías, en cuyo caso es deseable un test con particiones más finas. El contraste de Pearson en versión Tabla de Contingencia $m \times m$ (TC $m \times m$) teóricamente debe ser válido para cubrir tal situación, aunque es el test (P-T) derivado por Pesaran y Timmermann (1992), válido en el contexto de cambio direccional para particiones de m regiones, el que habitualmente se ha empleado en estos casos (por ejemplo, Greer (2003), con $m = 3$). Estos dos contrastes no son equivalentes asintóticamente en general, aunque sí lo son para $m = 2$.

Esencialmente, los tests mencionados se basan en comparar las frecuencias observadas de los datos y previsiones en cada región con las frecuencias teóricas bajo la hipótesis nula. La hipótesis de contraste es, en todos los casos, la no utilidad de las previsiones, aunque la formulación de la misma difiere ligeramente según el test. Merton (1981) y Henriksson-Merton (1981) definen las previsiones como no útiles si la distribución de probabilidad del usuario sobre la variable a predecir no se modifica al recibir las previsiones (test H-M), mientras Schnader y Stekler (1990) presentan la no utilidad de las predicciones en este contexto como independencia estocástica entre previsiones y datos (test TC 2×2), pero muestran que dicha hipótesis es equivalente a la de H-M. Además, en el caso de dos regiones en la partición, no solo estas dos hipótesis nulas son equivalentes sino que tienen una interpretación muy simple e intuitiva: las previsiones son útiles si y solo si son superiores a las generadas por un modelo "naive", que siempre predice el mismo signo/región (ver Stekler (1994)). La hipótesis nula implícita en el Test Binomial (B) utilizado por Greer (2003) es que las previsiones no son útiles si la probabilidad de éxito en la predicción no supera la que podría asociarse a la fortuna, 0,5. Respecto a los contrastes basados en particiones de m regiones, la hipótesis nula de ausencia de utilidad en las previsiones se identifica con independencia entre previsiones y datos en el caso del test TC $m \times m$ y con una variante de ésta última, menos restrictiva, en el contraste P-T.

Los contrastes mencionados presentan problemas obvios. Por un lado, aquellos que particionan el dominio de los datos solo en dos regiones únicamente valoran las previsiones en función de su capacidad de adelantar el signo de la variable objeto de predicción, pero ignoran totalmente la magnitud del error cometido (véase, por ejemplo, el comentario en Stekler (1994), pág.502). Por otro lado, los tests que utilizan particiones genéricas de m regiones identifican utilidad de previsiones con no independencia estocástica, enfoque claramente insatisfactorio. Sorprendentemente, estos contrastes se pronuncian sobre si las previsiones son "útiles" sin introducir en su especificación una función de utilidad o pérdida.[2] Esta crítica motiva nuestro trabajo, que trata de ser una solución de compromiso entre este tipo de contrastes de hipótesis (H-M, B, TC 2×2, TC $m \times m$, P-T) y las medidas cuantitativas habituales para valorar previsiones, tipo MSE o MAE.

[1] Las siglas entre paréntesis se refieren a la notación que usaremos para estos tests en lo sucesivo.

[2] Sospechamos que la causa de esa forma de proceder puede ser la dificultad de obtener un estadístico de contraste adecuado cuando se incorpora una función de pérdida continua estándar, o, en caso de tenerlo, la dificultad de derivar su distribución asintótica. Volveremos a este asunto a lo largo del capítulo.

Lógicamente, el planteamiento que nosotros sugerimos pasa por introducir una función de pérdida para valorar las previsiones, y, en base a ella, construir una hipótesis formal adecuada para contrastar la hipótesis subyacente de ausencia de utilidad de éstas. Lo peculiar de nuestra propuesta es la función de pérdida que sugerimos, que consistirá en particionar el dominio de los datos en m regiones y asignar pérdidas (utilidades, desde el punto de vista inverso) a cada uno de los m^2 "cuadrantes" en que podrían situarse el dato de un periodo y la previsión asociada, de modo que se valore no solo el acierto del signo de la variable a prever sino también la magnitud del error cometido. Denominaremos a dicha función como "discreta". Este tipo de criterio es muy interesante para muchas aplicaciones prácticas, del tipo a las mencionadas arriba.[3] Los beneficios a nivel conceptual de la función discreta se describieron en la Introducción a esta Tesis, y volveremos a ellos en el apartado 3.2.2. Pero, además, el hecho de emplear esta función va a facilitar enormemente la derivación de contrastes de hipótesis adecuados para el contexto que nos ocupa, por razones que se verán a lo largo del trabajo.

En el trabajo a continuación, se proponen una serie de tests de hipótesis para contrastar la utilidad de las previsiones, en el marco expuesto, que consideramos más apropiado. Se llevan a cabo ejercicios de Monte Carlo, para determinar propiedades estadísticas de dichos contrastes y para comprobar su diferente comportamiento respecto a los cinco tests de la literatura de análisis direccional citados arriba.

Todo el trabajo se realiza bajo el supuesto de inexistencia de incertidumbre paramétrica asociada a las previsiones. Cuando éstas proceden de modelos estimados, existe una nueva fuente de variabilidad en las previsiones, debida a la estimación de los parámetros desconocidos de los modelos predictivos. En teoría, dicha variabilidad debe ser tenida en cuenta en el diseño de cualquier contraste que evalúa un conjunto de previsiones generadas de esa manera. No obstante, ignoraremos este asunto a la hora de derivar nuestros tests (igual que ha ocurrido con los contrastes estándar de la literatura, mencionados arriba). En el Capítulo 3 de esta Tesis se volverá a esta cuestión, presentándose evidencia teórica y empírica a favor de la robustez de nuestros tests a la presencia de incertidumbre paramétrica,[4] de modo que las conclusiones de este capítulo pueden considerarse válidas incluso en el caso de que las previsiones se hayan generado con modelos estimados.

El capítulo se organiza como sigue: en la Sección 2 se exponen formalmente los tests utilizados habitualmente en la literatura sobre el tema que nos ocupa. La Sección 3 resume las críticas que encontramos a dichos tests y ofrece nuestro planteamiento alternativo, mientras en la sección posterior se presentan detalladamente los contrastes que proponemos. En la Sección 5 explicamos el análisis de Monte Carlo utilizado para evaluar comparativamente nuestros tests y los mencionados en la Sección 2, y los resultados obtenidos al respecto. Finalmente, la Sección 6 cita las conclusiones fundamentales que se pueden inferir del trabajo.

[3] En ese tipo de series, el acierto en la previsión del signo es esencial, pero, además, interesa tener en cuenta la cuantía del error cometido. Por otro lado, el hecho de que el número de regiones de la partición sea finito (normalmente pequeño) es un punto de partida perfectamente asumible para estas series, para las que no exigir alta precisión cuantitativa en la previsión parece razonable. Por último, este planteamiento evitará el efecto distorsionador de los valores atípicos, que no son infrecuentes en estas variables.

[4] En realidad, de los cuatro tests que presentamos en este documento para evaluar utilidad de previsiones, los resultados del capítulo 3 garantizan que la afirmación realizada aplica precisamente a los dos tests que recomendaremos (por otras razones) al usuario.

2. Métodos estadísticos utilizados en la literatura

A lo largo de toda la sección se designarán por y_t y v_t el dato y previsión en/para t, respectivamente, y por T el tamaño muestral.

Los tres contrastes a continuación (H-M, TC 2×2 y B) utilizan una partición del dominio de y_t, v_t de dos regiones, que designaremos por r_1, r_2. En el caso del Test Binomial, estas regiones se corresponderán forzosamente con los valores positivos y negativos de los datos, respectivamente. En los otros dos contrastes, no tiene porqué ser así, aunque ésta será el caso más habitual en la práctica (Schnader y Stekler (1990) representa una excepción).

2.1. Test H-M

Sean $P_1 = P(v_t \in r_1 | y_t \in r_1)$, $P_2 = P(v_t \in r_2 | y_t \in r_2)$, Henriksson y Merton demostraron que:

1) $f^1(x) = f^2(x)$ si y solo si $P_1 + P_2 = 1$, siendo $f^1(x)$ la distribución de probabilidad a priori (antes de conocer las previsiones) del usuario respecto a los datos x y $f^2(x)$ la distribución a posteriori (una vez se reciben las previsiones).

2) Bajo la hipótesis $P_1 + P_2 = 1$, el número de previsiones correctas del tipo $v_t \in r_1$, que se designará por X, sigue una distribución Hipergeométrica con parámetros (T, n_1, m), siendo n_1 el número de casos $y_t \in r_1$ y m el número de casos $v_t \in r_1$.

En base a 1) y 2) proponen el contraste siguiente (H-M):

a) Hipótesis: La hipótesis nula (las previsiones no son útiles) será $H_0 = P_1 + P_2 = 1$ (esto se deduce inmediatamente de la definición de utilidad de Henriksson y Merton (1981) -las previsiones no son útiles si la distribución de probabilidad a priori del usuario no se modifica al recibir las previsiones- y del resultado expuesto en 1)), y la alternativa será $H_1 = P_1 + P_2 > 1$, por lo que el contraste es unilateral.

b) Estadístico de contraste y distribución exacta bajo H_0:

$X \sim$Hipergeométrica (T, n_1, m): $P(X = x) = \binom{T}{m}^{-1}\binom{n_1}{x}\binom{T-n_1}{m-x}$.

c) Implementación del test: la región crítica es la cola superior de dicha distribución. Sea X_0 el valor observado de X y sea α el nivel de significación prefijado, se rechaza la hipótesis nula si $X_0 \geq \lambda_\alpha$, siendo λ_α el valor para el que $P(X \geq \lambda_\alpha) = \alpha$ en una distribución Hipergeométrica de parámetros (T, n_1, m).[5]

Aunque no entraremos a demostrar que, bajo el supuesto de previsiones no útiles, X sigue una distribución Hipergeométrica de parámetros (T, n_1, m), dicho resultado es más o menos intuitivo. Imaginemos que los datos $y_1, ..., y_T$ se presentan como un conjunto de T bolas numeradas 1 a T que contienen una nota con el tipo de dato que resultó, tipo r_1 o tipo r_2, siendo la nota no visible para el predictor. Éste debe pegar una etiqueta negra en el exterior de cada bola i para la que predijo que el dato asociado y_i era de tipo r_1. Hizo m previsiones de tipo r_1, así que pondrá m etiquetas negras, en total. El número de bolas que, realmente, contiene el tipo r_1 es n_1. ¿Cuál es la probabilidad de haber pegado x etiquetas negras correctas? Si las previsiones no son útiles, digamos que el acierto será aleatorio, de modo que dicha probabilidad viene dada por el número de casos favorables sobre casos posibles. Se producen x aciertos y $m - x$ fallos si las x primeras etiquetas negras son correctas y el resto erróneas, suceso cuya probabilidad será $\frac{n_1}{T}\frac{n_1-1}{T-1}...\frac{n_1-x+1}{T-x+1}\frac{T-n_1}{T-x}\frac{T-n_1-1}{T-x-1}...\frac{T-n_1-(m-x)+1}{T-m+1}$, y en cualquier otra permutación con x aciertos. Por tanto, la probabilidad de que se hayan pegado x etiquetas negras correctas es exactamente $\binom{T}{m}^{-1}\binom{n_1}{x}\binom{T-n_1}{m-x}$, es decir, X sigue una distribución Hipergeométrica de parámetros (T, n_1, m). Obviamente, cuanto más útiles sean las previsiones, mayor tenderá a ser el valor observado X_0, y menos probables serán en la distribución citada valores iguales o mayores que X_0, por lo que la región de rechazo es la cola superior. Véase que H-M está considerando implícitamente que n_1 y m son parámetros, y no variables aleatorias.

Antes de terminar la exposición del test H-M, presentamos una sencilla interpretación de su hipótesis nula, apuntada por Schnader y Stekler (1990): supóngase un modelo/método de previsión "naive", que hiciera predicciones siempre del mismo tipo, o bien $v_t \in r_1$, en cuyo caso $P_1 = 1$ y $P_2 = 0$ siempre, o bien $v_t \in r_2$, en cuyo caso $P_1 = 0$ y $P_2 = 1$ siempre. Así, un método de previsión naive de este tipo siempre verificará $P_1 + P_2 = 1$, y, de este modo, contrastar $H_0 = P_1 + P_2 = 1$ puede interpretarse como contrastar si las previsiones son superiores a las de un modelo naive, que las sitúe siempre en la misma región.

[5] En realidad, no habrá ningún valor que verifique tal igualdad (salvo por casualidad). Por ello, en teoría, debería aplicarse una versión aleatorizada del test. Este asunto se trata en el Apéndice A. No obstante, para la exposición que aquí sigue, este detalle puede ignorarse.

2.2. Test TC 2×2

Schnader y Stekler (1990) proponen utilizar un test de Pearson respecto a una Tabla de Contingencia 2×2 para contrastar utilidad de previsiones, definida como dependencia estocástica entre éstas y los datos. El test H-M y el test de Schnader y Stekler (TC 2×2) son asintóticamente equivalentes, pero difieren en muestras finitas, siendo H-M más potente (Ash, Smyth y Heravi (1998), pág. 385). La discrepancia se debe a que H-M considera n_1 y m como parámetros en vez de como variables aleatorias, de modo que está tratando las frecuencias de las filas y columnas de la Tabla de Contingencia implícita como si fueran conocidas, lo contrario que TC 2×2 (Pesaran y Timmermann (1992), pág. 462). El contraste propuesto es el siguiente:

a) Hipótesis: La hipótesis nula (las previsiones no son útiles) es $H_0 = p_{ij} = p_i^y p_j^v$, $i = 1, 2$, $j = 1, 2$, siendo p_{ij} la probabilidad del suceso $\{y_t \in r_i, v_t \in r_j\}$, mientras p_i^y (p_j^v) es la probabilidad marginal del suceso $\{y_t \in r_i\}$ ($\{v_t \in r_j\}$). Pese a las aparentes discrepancias entre esta hipótesis y la del test H-M, ambas están muy relacionadas, ya que $P_1 = P(v_t \in r_1 | y_t \in r_1) = p_{11}/p_1^y$, $P_2 = P(v_t \in r_2 | y_t \in r_2) = p_{22}/p_2^y$ y, bajo la hipótesis de independencia entre previsiones y datos, se tiene que $P_1 + P_2 = p_1^v + p_2^v = 1$. La hipótesis alternativa es $H_1 \equiv \neg H_0$.

b) Estadístico de contraste y distribución asintótica bajo H_0:

$S_{TC2} = \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{(n_{ij} - T\widehat{p}_i^y \widehat{p}_j^v - 1/2)^2}{T\widehat{p}_i^y \widehat{p}_j^v} \xrightarrow{L} \chi_1^2$, siendo n_{ij} la frecuencia asociada al suceso $\{y_t \in r_i, v_t \in r_j\}$, mientras $\widehat{p}_i^y = \frac{\sum_{j=1}^{2} n_{ij}}{T}$ y $\widehat{p}_j^v = \frac{\sum_{i=1}^{2} n_{ij}}{T}$. Véase que S_{TC2} es el estadístico de Pearson, pero introduciendo la corrección de Yates, adecuada en Tablas de Contingencia 2×2, al haber un solo grado de libertad (corrección recomendada, por ejemplo, en Schnader y Stekler (1990), nota a pie de página número 7, pág.104).

c) Implementación del test: La región crítica del contraste es la cola superior de la distribución anterior. Por tanto, sea S_{TC2}^0 el valor observado de S_{TC2} y sea α el nivel de significación prefijado para el test, se rechaza H_0 si $S_{TC2}^0 > \lambda_\alpha$, siendo λ_α el valor para el que $P(S_{TC2} > \lambda_\alpha) = \alpha$ en una distribución χ_1^2.

En principio, el contraste no podría llevarse a cabo si existe algún i tal que $\widehat{p}_i^y = 0$ ó/y algún j tal que $\widehat{p}_j^v = 0$. En la práctica, podemos evitar este obstáculo procediendo de forma sencilla tal y como se explica en el apartado 4.6.

2.3. Test Binomial

Greer (2003) utiliza un test estándar sobre la distribución Binomial para contrastar utilidad de previsiones, definida ésta en términos de su capacidad de anticipar el signo que tomarán los datos:

Sea la variable indicatriz d_t tal que $d_t = 1$, si la previsión fue correcta en signo y $d_t = 0$, si no. Designemos por θ la probabilidad teórica de que la previsión sea correcta en cuanto a signo, es decir, $\theta = P(d_t = 1)$.

a) Hipótesis: La hipótesis nula es $H_0 = \theta = 0,5$. Es decir, la hipótesis nula es que las previsiones tienen el mismo ratio de acierto respecto al signo que el que podría obtenerse por fortuna. Obviamente, la hipótesis alternativa será $H_1 = \theta > 0,5$.

b) Estadístico de contraste y distribución exacta bajo H_0:

$S_B = \sum_{t=1}^{T} d_t \sim B(T, \frac{1}{2})$. Para muestras grandes se puede usar la versión asintótica: $\frac{S_B - 0,5T}{\sqrt{T0,25}} \xrightarrow{L} N(0, 1)$.

c) Implementación del test: se trata de un contraste unilateral, siendo de nuevo la región crítica la cola superior. Sea S_B^0 el valor observado del estadístico S_B y α el nivel de significación prefijado, se rechaza la hipótesis nula si $S_B^0 \geq \lambda_\alpha$, siendo λ_α el valor para el $P(S_B \geq \lambda_\alpha) = \alpha$ en una distribución Binomial de parámetros $(T, \frac{1}{2})$.[6]

Debemos apuntar que, dado que la derivación de la distribución del estadístico S_B es conocida solo bajo el supuesto de que las variables $d_1, ..., d_T$ sean mutuamente independientes (además de seguir la misma distribución), es aconsejable, previa a la ejecución del test Binomial, llevar a cabo un contraste de

[6] En realidad, no habrá ningún valor que verifique tal igualdad (salvo por casualidad). Por ello, en teoría, debería aplicarse una versión aleatorizada del test. Este asunto se trata en el Apéndice A. No obstante, para la exposición que aquí sigue, este detalle puede ignorarse.

aleatoriedad para verificar si $d_1, ..., d_T$ constituyen una muestra aleatoria, por ejemplo, un test de rachas. Greer (2003) implementa el test Binomial usando este paso previo.[7]

A continuación se presentan los tests estándar de la literatura para el caso de particiones de m regiones.

Se utilizará la notación n_{ij} para designar la frecuencia del suceso $\{y_t \in r_i, v_t \in r_j\}$, mientras la notación p_{ij}, p_i^y, p_j^v designará las probabilidades siguientes: $p_{ij} = P(y_t \in r_i, v_t \in r_j)$, $p_i^y = P(y_t \in r_i)$, $p_j^v = P(v_t \in r_j)$, para $i = 1, ..., m$ y $j = 1, ...m$. Por otro lado, $\widehat{p}_{ij}$, $\widehat{p}_i^y$, $\widehat{p}_j^v$ simbolizarán las estimaciones de máxima verosimilitud (MV) de las probabilidades teóricas anteriores, es decir, $\widehat{p}_{ij} = \frac{n_{ij}}{T}$, $\widehat{p}_i^y = \frac{\sum_{j=1}^{m} n_{ij}}{T}$ y $\widehat{p}_j^v = \frac{\sum_{i=1}^{m} n_{ij}}{T}$.

2.4. Test TC $m \times m$

Aunque este test apenas ha sido empleado explícitamente en la literatura,[8] es una extensión inmediata del presentado en la subsección 2.2 y podría utilizarse para contrastar utilidad de previsiones (en el sentido de dependencia estocástica entre datos y previsiones) en un contexto de m regiones. Se trata del conocido contraste no paramétrico Chicuadrado sobre independencia entre dos variables. De él existen dos versiones, la obtenida directamente a partir de la implementación asintótica del test de Razón de Verosimilitudes, y la versión de Pearson, que es la más popular y la que exponemos aquí:

La hipótesis nula será $H_0 = p_{ij} = p_i^y p_j^v$, $i = 1, 2, ..., m$, $j = 1, 2, ..., m$ y la alternativa, $H_1 \equiv \neg H_0$. El estadístico de contraste bajo H_0 es $S_{TCm} = \sum_{i=1}^{m} \sum_{j=1}^{m} \frac{(n_{ij} - T\widehat{p}_i^y \widehat{p}_j^v)^2}{T\widehat{p}_i^y \widehat{p}_j^v}$ y su distribución límite es $\chi^2_{(m-1)^2}$. La región crítica del contraste es la cola superior de la distribución anterior. Por tanto, sea S^0_{TCm} el valor observado de S_{TCm} y sea α el nivel de significación prefijado, se rechaza H_0 si $S^0_{TCm} > \lambda_\alpha$, siendo λ_α el valor para el que $P(S_{TCm} > \lambda_\alpha) = \alpha$ en una distribución $\chi^2_{(m-1)^2}$.

En principio, el contraste no podría llevarse a cabo si existe algún i tal que $\widehat{p}_i^y = 0$ ó/y algún j tal que $\widehat{p}_j^v = 0$. En la práctica, podemos evitar este obstáculo procediendo de forma sencilla tal y como se explica en el apartado 4.6.

2.5. Test P-T

Pesaran y Timmermann (1992) presentaron un test para contrastar utilidad de previsiones, en base al concepto de dependencia estocástica entre previsiones y datos, en el contexto de análisis direccional y partición de $m \geq 2$ regiones. Los autores afirman que su contraste es asintóticamente equivalente a la Tabla de Contingencia cuando $m = 2$, pero muestran que dicha equivalencia no se verifica para valores superiores de m. El test P-T es el siguiente:

a) Hipótesis: La hipótesis nula es $H_0 = \sum_{i=1}^{m} p_{ii} = \sum_{i=1}^{m} p_i^y p_i^v$. Esta hipótesis no coincide con la hipótesis de independencia estocástica entre y_t, v_t, sino que es menos restrictiva, ya que exige el cumplimiento de la condición $p_{ij} = p_i^y p_j^v$ solo en los casos $i = j$, y únicamente de manera agregada. De tal modo, es obvio que la hipótesis de independencia implica ésta, mientras lo contrario no es cierto. Solo en el caso $m = 2$ ambas son equivalentes, afirmación que demostramos en el Apéndice B. No obstante, pese a que la hipótesis nula de P-T sea menos restrictiva que la de TC $m \times m$, Pesaran y Timmermann (1992) constatan que, en general, el test TC $m \times m$ es más conservador − en el sentido de menor probabilidad de rechazo de la hipótesis nula − que el suyo, y presentan ciertas pruebas analíticas sobre este particular. Como podrá comprobarse en secciones posteriores, la evidencia empírica ofrecida por nuestros ejercicios de simulación confirman esta apreciación.

[7] En realidad, sería conveniente proceder de manera análoga en todos los tests que se estudian en este documento. Volveremos a este asunto en el apartado 4.6.

[8] Una excepción es Pesaran y Timmermann (1992), donde los autores presentan una aplicación práctica, con $m = 3$, en la que comparan los resultados de su contraste (P-T) con los de una Tabla de Contingencia 3×3.

b) Estadístico de contraste y distribución asintótica bajo H_0:

$S_{PT} = \sqrt{T} S_n \widehat{W}^{-1/2} \xrightarrow{L} N(0,1)$, siendo $S_n = h(\widehat{P})$, $P = (p_{11}, p_{12}, ..., p_{1m}, p_{21}, ..., p_{2m}, ..., p_{m1}, ..., p_{mm})'$ y $\widehat{P} = (\frac{n_{11}}{T}, \frac{n_{12}}{T}, ..., \frac{n_{m-1m}}{T}, \frac{n_{mm}}{T})$ la estimación MV para P, mientras h es una función de R^{m^2} en R cuya expresión es $h(P) = \sum_{i=1}^{m} p_{ii} - \sum_{i=1}^{m} p_i^y p_i^v$. Por lo tanto, $S_n = \sum_{i=1}^{m} \widehat{p}_{ii} - \sum_{i=1}^{m} \widehat{p}_i^y \widehat{p}_i^v$. Finalmente, $\widehat{W} = \left(\frac{\partial h}{\partial P}\right)_{P=\widehat{P}} V_P \left(\frac{\partial h}{\partial P}\right)'_{P=\widehat{P}}$, siendo $V_P = (\widehat{\Psi} - \widehat{P}\widehat{P}')$, donde $\widehat{\Psi}$ es una matriz diagonal cuyos elementos en la diagonal son los elementos de $\widehat{P}$.

c) Implementación del test: El cálculo de S_n, $\widehat{\Psi}$, $\widehat{W}$ solo depende de $\widehat{P}$ y $\widehat{p}_i^y, \widehat{p}_j^v$. El contraste es bilateral. Sea S_{PT}^0 el valor observado de S_{PT} y sea α el nivel de significación prefijado, se rechaza H_0 si $\left|S_{PT}^0\right| > \lambda_\alpha$, siendo λ_α el valor para el que $\Phi(\lambda_\alpha) = 1 - \alpha/2$, donde Φ denota la función de distribución $N(0,1)$.

La derivación del test P-T se obtiene de la aplicación de los teoremas a continuación, todos ellos bien conocidos:

Teorema 1: Sea $\widehat{\theta}_{MV}$ el estimador MV de un vector paramétrico θ de dimensión k, su distribución límite es: $\sqrt{T}(\widehat{\theta}_{MV} - \theta) \xrightarrow{L} N(0_k, I_1^{-1}(\theta))$, siendo $I_1(\theta)$ la matriz de información de Fisher para una muestra de longitud uno.

Teorema 2: Sea $(N_1, ..., N_k)$ un vector de frecuencias cuya distribución es Multinomial de parámetros (T, p), con $p = (p_1, ..., p_k)'$. El estimador MV del vector p es $\widehat{p} = (N_1/T, ..., N_k/T)'$. Además, la matriz de información asociada a ese caso para una muestra unitaria es $I_1(p) = (A - pp')^{-1}$, siendo A una matriz diagonal con los elementos de p en dicha diagonal.

Teorema 3: Sea $\sqrt{T}(Z_T - \theta) \xrightarrow{L} N(0_k, \Sigma)$ y sea g una función diferenciable de R^k en R^b, entonces $\sqrt{T}(g(Z_T) - g(\theta)) \xrightarrow{L} N(0, \Gamma)$, siendo $\Gamma = \nabla g(\theta) \Sigma \nabla g(\theta)'$, siendo $\nabla g(\theta)$ la matriz $b \times k$ gradiente de la función vectorial g.

De los Teoremas 1 a 3 se deduce que $\sqrt{T} S_n W^{-1/2} \xrightarrow{L} N(0,1)$.

Teorema 4: (i) Si $X_T \xrightarrow{L} X$, $Y_T \xrightarrow{p} c$, entonces $X_T Y_T \xrightarrow{L} cX$; (ii) $X_T \xrightarrow{p} c$ y $g(.)$ es continua, entonces $g(X_T) \xrightarrow{p} g(c)$; (iii) $\widehat{p} \xrightarrow{p} p$.

Por tanto, $S_{PT} = \sqrt{T} S_n \widehat{W}^{-1/2} \xrightarrow{L} N(0,1)$, q.e.d.

2.6. Nota sobre el uso de tests estándar para contrastar utilidad de las previsiones

Una dificultad esencial para contrastar si un conjunto de previsiones son o no útiles para predecir una variable es la especificación de las hipótesis formales que "traduzcan" tal idea. Como se ha visto en los apartados anteriores, una posibilidad consiste en contrastar si datos y previsiones son estocásticamente independientes (previsiones no útiles). Para responder a dicha pregunta, existen tests estándar bien conocidos. La primera opción sería usar un contraste paramétrico sobre el coeficiente de correlación lineal ρ_{yv} a través de su análogo muestral $\widehat{\rho}_{yv}$, pero su distribución solo es conocida si la muestra aleatoria simple bidimensional $\{(y_1, v_1), ..., (y_T, v_T)\}$ procediera de una distribución Normal Bivariante, supuesto que, obviamente, no puede asumirse. Eliminada esta opción, se puede acudir a los contrastes no paramétricos sobre independencia entre dos variables. Uno de ellos es el test Chicuadrado (bien en la versión de Razón de Verosimilitudes o bien en la versión de Pearson, que es la usada habitualmente), que es exactamente el que hemos denotado por TC (para los casos $m = 2$ y $m \geq 2$). Además de éste, existen otros dos tests de independencia no paramétricos estándar: el contraste τ de Kendall y el contraste de correlación entre rangos, de Spearman. Son válidos en muestras finitas, pero su derivación solo es correcta si la distribución bidimensional asociada a la muestra aleatoria simple $\{(y_1, v_1), ..., (y_T, v_T)\}$ es continua. Además, tienen el problema de que su distribución exacta no es estándar, sino que está tabulada. Quizá por ello, no suelen emplearse en este contexto, pero serían una alternativa válida.

Por su parte, el test Binomial no contrasta independencia estocástica entre datos y previsiones. Se trata de un test paramétrico que contrasta el valor del parámetro θ de la distribución de Bernouilli asociada a la variable aleatoria d, variable que asigna, con probabilidad θ, valor 1 si la previsión tiene el signo del dato, y 0, si no.

Los otros dos tests, H-M (válido en muestras finitas) y P-T (contraste de significación de validez asintótica) no son estándar en la literatura estadística general, sino que han sido desarrollados específicamente

para aplicarse en el contexto de evaluación de capacidad predictiva de un conjunto de previsiones. Ninguno de los dos contrasta la no utilidad de las previsiones exactamente a través de su independencia estocástica con los datos y previsiones, pero su hipótesis nula sí está muy relacionada con ésta.

Salvo TC 2×2, los contrastes de dos regiones tienen la ventaja de ser válidos para muestras finitas y no rechazar la hipótesis de ausencia de utilidad de las previsiones si la correlación entre datos y previsiones es negativa, al ser su hipótesis alternativa unilateral. En cambio, TC y P-T, cuyas hipótesis alternativas son bilaterales, presentan el problema citado, como veremos enseguida, además de ser válidos solo para muestras grandes. Por el contrario, permiten una partición más fina que la de los anteriores, los cuales solo valoran una previsión por su capacidad de predecir correctamente el signo del dato, criterio que, a priori, puede resultar impreciso. Estas cuestiones sirven para motivar la sección a continuación.

3. Motivación del trabajo. Enfoque propuesto

3.1. Crítica a los contrastes de la literatura

Los tests anteriores están siendo utilizados habitualmente en la literatura para evaluar la utilidad de un conjunto de previsiones respecto de los cambios en una serie económica del tipo de las mencionadas en la sección 1. En muchas ocasiones, la evaluación de las previsiones será incompleta si se realiza exclusivamente en base al acierto o fallo respecto del signo del cambio, es decir, si se emplea alguno de los tres primeros tests de la sección anterior.

En cambio, los tests TC $m \times m$ y P-T utilizan una partición del dominio de los datos tan fina como se quiera, pero pueden cometer importantes errores en la valoración de las previsiones, debido a la definición de la hipótesis nula y, sobre todo, de la hipótesis alternativa, que se establece simplemente como negación de la nula. Es inmediato comprobar que para estos contrastes no es relevante la distancia entre la región en que se situó el dato y la prevista, algo que no parece razonable. El motivo que suscita este problema es que los tests contrastan la hipótesis de independencia estocástica entre previsiones y datos y la negación de esta hipótesis constituye la hipótesis alternativa, criterio éste que resulta poco satisfactorio para definir previsiones útiles. Si las previsiones son independientes estocásticamente de los datos, no son de utilidad para el usuario, indudablemente. Pero lo contrario no tiene porqué ser cierto.

Un ejemplo muy simple basta para ilustrar estas ideas. Supongamos que para contrastar la no utilidad de las previsiones disponibles respecto de los cambios de una serie determinada se utiliza una partición del dominio de los datos en cuatro regiones (r_1, r_2, r_3, r_4): G-, P-, P+, G+, donde el símbolo + ó - indica el signo de los datos, y la letra su magnitud en valor absoluto (ie, "negativo grande/pequeño", "positivo grande/pequeño"). Se dispone de una muestra de 100 datos sobre los cambios de la serie y las previsiones correspondientes, generadas por tres modelos de predicción alternativos. Para las previsiones se aplica la misma partición que para los datos. Supongamos que las observaciones muestrales y las previsiones generadas por los modelos pueden sintetizarse en matrices, cuyo elemento (i, j) representa el número de veces que el dato y_t y la previsión v_t se situaron en el cuadrante (i, j). Las matrices a continuación siguen dicha representación y se corresponden con los resultados de los tres modelos predictivos:

$$M_1 = \begin{array}{cc} & \\ y_t & \end{array} \begin{array}{c} \\ \text{G-} \\ \text{P-} \\ \text{P+} \\ \text{G+} \end{array} \begin{array}{c} v_t \\ \begin{array}{cccc} \text{G-} & \text{P-} & \text{P+} & \text{G+} \end{array} \\ \begin{array}{cccc} \mathbf{0} & \mathbf{0} & 0 & \textit{25} \\ \mathbf{0} & \mathbf{0} & 25 & \textit{0} \\ \textit{0} & 25 & \mathbf{0} & \mathbf{0} \\ \textit{25} & 0 & \mathbf{0} & \mathbf{0} \end{array} \end{array}$$

$$M_2 = \begin{array}{cc} & \\ y_t & \end{array} \begin{array}{c} \\ \text{G-} \\ \text{P-} \\ \text{P+} \\ \text{G+} \end{array} \begin{array}{c} v_t \\ \begin{array}{cccc} \text{G-} & \text{P-} & \text{P+} & \text{G+} \end{array} \\ \begin{array}{cccc} \mathbf{0} & \mathbf{25} & 0 & \textit{0} \\ \mathbf{25} & \mathbf{0} & 0 & \textit{0} \\ \textit{0} & 0 & \mathbf{0} & \mathbf{25} \\ \textit{0} & 0 & \mathbf{25} & \mathbf{0} \end{array} \end{array} \qquad M_3 = \begin{array}{cc} & \\ y_t & \end{array} \begin{array}{c} \\ \text{G-} \\ \text{P-} \\ \text{P+} \\ \text{G+} \end{array} \begin{array}{c} v_t \\ \begin{array}{cccc} \text{G-} & \text{P-} & \text{P+} & \text{G+} \end{array} \\ \begin{array}{cccc} \mathbf{25} & \mathbf{0} & 0 & \textit{0} \\ \mathbf{0} & \mathbf{25} & 0 & \textit{0} \\ \textit{0} & 0 & \mathbf{25} & \mathbf{0} \\ \textit{0} & 0 & \mathbf{0} & \mathbf{25} \end{array} \end{array}$$

Como puede verse, tanto los datos como las previsiones se situaron 25 veces en cada una de las cuatro regiones, pero la disposición en los cuadrantes es muy distinta según el modelo. Aunque no hemos definido ninguna función de pérdida todavía, parece razonable que se desee que las previsiones sean certeras en signo y lo más precisas posible, en el sentido de que la distancia entre la región en que se sitúen la previsión sea próxima a aquella en la que lo hizo el dato. Hemos destacado en negrita los cuadrantes en los que las previsiones serían certeras en signo, y en cursiva, aquellos cuadrantes en los que la previsión, además de errónea en signo, sería lo menos precisa posible. Por ejemplo, interpretemos la primera fila de cualquier matriz. En ella se encuentran todos los casos en que los datos fueron grandes y negativos (G-). Bien, si las previsiones se sitúan en una de las dos primeras regiones (G-, P-), al menos se predijo correctamente el signo (cuadrantes (1,1) y (1,2)), y con precisión total si se situaron en concreto en la región G- (cuadrante (1,1)). Los cuadrantes (1,3) y (1,4) implican errores en signo, siendo la peor previsión posible la asociada a la región G+ (cuadrante (1,4)), dado que la imprecisión es máxima (la previsión se encuentra en la región más lejana posible a la del dato). Análogamente, se interpretan el resto de las filas. Ahora ya estamos en condiciones de entender los resultados de cada modelo predictivo.

El Modelo 1 (M1) falla en la previsión del signo siempre, incurriendo, además, en máxima imprecisión el 50 % de las veces (cuadrantes (1,4) y (4,1)). En cambio, el Modelo 2 (M2) prevé correctamente el signo siempre, aunque nunca lo hace con máxima precisión (ya que sus predicciones no se sitúan nunca en la diagonal). Por su parte, el Modelo 3 (M3) es totalmente certero siempre, tanto en signo como en magnitud.

Sin embargo, el valor del estadístico del contraste TC $m \times m$ resulta $S_{TCm} = 292,31$ en los tres casos, mientras el estadístico del test P-T toma valores $S_{PT} = -353,55$, $S_{PT} = -353,55$ y $S_{PT} = 74,94$, para los modelos 1, 2 y 3, respectivamente.[9] Aunque las previsiones de los tres modelos no sean independientes estocásticamente de los datos, es obvio que las generadas por M1 no tienen ninguna valía para el usuario, mientras las de M3 son óptimas y las de M2 pueden resultar útiles, aunque no sean totalmente precisas. Lo razonable sería que los contrastes no rechazaran la hipótesis de no utilidad de las previsiones en el caso de M1, la rechazaran en el caso de M3, mientras el resultado deseable para el contraste respecto de las previsiones de M2 sería discutible, y dependería de juicios de valor del usuario en la aplicación concreta en que se encuentre (precisamente, la necesidad de introducir estos juicios de valor en los tests, a través de una función de pérdida, es lo que motiva el planteamiento de los contrastes que propondremos en próximas secciones). Sin embargo, no solo los dos tests llevan al rechazo de la hipótesis nula de no utilidad de las previsiones, sino que incluso los valores de los estadísticos de contraste son idénticos para las previsiones de los modelos 1 y 2 en el test P-T y para las previsiones de los tres modelos en el caso del contraste TC $m \times m$.

3.2. Planteamiento propuesto: la función de pérdida discreta

3.2.1. Enfoque general

Por todo esto, definimos un planteamiento distinto a partir del que se construirán nuestros tests. En primer lugar, utilizaremos particiones genéricas de m regiones, con $m > 2$, para evitar el problema mencionado de considerar solo el signo de la previsión y no la magnitud del error en la valoración de las previsiones. Y en segundo lugar, introduciremos una función de pérdida, que asigne una penalización numérica a cada uno de los "cuadrantes" o pares de regiones (r_i, r_j) en que pueden situarse el dato y la previsión correspondiente, tipo de función que denominaremos "discreta". El uso de esta función de pérdida permitirá establecer las hipótesis nula y alternativa de forma que el concepto de utilidad y no utilidad de las previsiones, que es lo que se pretende contrastar, quede definido en cierta parte por el usuario y responda a interpretaciones más razonables y realistas que la dependencia e independencia estocástica, valorándose tanto el signo de la previsión como la magnitud del error cometido.

En el contexto en el que se está pensando a la hora de diseñar nuestros tests (cambio direccional en series económicas como las mencionadas en la Sección 1), lo razonable es que las m regiones que definan la partición se establezcan simétricamente en torno al valor cero, en principio, de modo que la mitad se correspondan a valores positivos, distinguiendo por magnitudes, y la otra mitad, a valores negativos. No obstante, otras decisiones son perfectamente válidas. La elección sobre el número de regiones en la partición deberá tener en cuenta también el tamaño muestral. Una función de pérdida típica que utilizaremos en los experimentos de Monte Carlo asigna valores de un conjunto A formado por m elementos de un modo tal que, para cada región r_i en que se podría situar un dato, cada uno de los valores de A está siendo asignado a alguno de los m cuadrantes en que puede situarse, simultáneamente, la previsión. Sin embargo, muchos otras funciones alternativas son razonables. Ilustraremos estas ideas con un ejemplo sencillo de función de pérdida, precisamente del tipo de las que acabamos de mencionar:

[9] En realidad, los tests se han ejecutado para matrices ligeramente distintas (denótense por M_1', M_2' y M_3') a las escritas en el texto. En concreto, M_1', M_2' y M_3' son exactamente iguales a M_1, M_2 y M_3 solo que en la primera fila de cada matriz, se ha sustituido el valor 25 del único elemento no nulo por el valor 24, mientras el elemento (1,3) se ha rellenado con 1. La explicación es que el test P-T no puede ejecutarse en casos donde todas las frecuencias se encuentran en la diagonal de la matriz o bien todas fuera de ella, porque la varianza $\widehat{W}$ utilizada en su estadístico de contraste sería nula. Podríamos haber presentado en el texto directamente las matrices M_1', M_2' y M_3', pero creemos que la idea que el ejemplo pretende transmitir queda mucho más nítida usando las primeras.

$$
\begin{array}{cc|cccc}
 & & \multicolumn{4}{c}{v_t} \\
 & & (-\infty,-a) & (-a,0) & (0,+a) & (+a,+\infty) \\
\hline
 & (-\infty,-a) & 0 & 1 & 2 & 3 \\
y_t & (-a,0) & 1 & 0 & 2 & 3 \\
 & (0,+a) & 3 & 2 & 0 & 1 \\
 & (+a,+\infty) & 3 & 2 & 1 & 0
\end{array}, \tag{1}
$$

siendo a un valor real razonable para la aplicación concreta de que se trate.

Los conceptos de no utilidad (hipótesis nula) y utilidad (hipótesis alternativa) se establecerán en términos de probabilidades de que las previsiones incurran en las distintas pérdidas que la función de pérdida asigna. Así, los contrastes que sugeriremos más adelante también se basan en comparar frecuencias teóricas y observadas, pero dichas frecuencias ya no se corresponden directamente con las de datos y previsiones, sino con las de las pérdidas.

3.2.2. Ventajas de la función de pérdida discreta

Como se ha dicho arriba, el uso de funciones de pérdidas permite definir el concepto de utilidad de las previsiones de forma más razonable y menos rígida que a través de la mera no dependencia estocástica entre datos y previsiones. Pero, ¿por qué una función "discreta"? ¿qué ventajas tiene en este contexto?. Esencialmente, tres:

a) En primer lugar, este tipo de función es muy adecuada para muchas aplicaciones para las que la predicción correcta del signo es clave, mientras la magnitud del error de previsión es importante, pero basta distinguir entre magnitudes a grosso modo (grandes, intermedias y pequeñas, etc). Por ejemplo, esta es la situación si se pretende prever variables cualitativas, pero también es un contexto apropiado para valorar la previsión de muchas variables macroeconómicas o incluso financieras.

b) El usuario es quien decide la partición y asigna las pérdidas, lo que proporciona una gran flexibilidad, en vez de tener que ceñirse a una función estándar.

Estas dos ventajas conceptuales son intrínsecas de la función que proponemos (ya se mencionaron en la Introducción de la Tesis), pero, además, existe otra ventaja técnica muy relevante que afecta al marco que nos ocupa en este capítulo, el diseño de tests para contrastar utilidad de *un* conjunto de previsiones:

c) El uso de una función discreta facilita diseñar estadísticos de contraste con distribución asintótica conocida, algo que podría no estar garantizado si se utilizara una función de pérdida continua. En primer lugar, porque no sería fácil definir el estadístico de contraste apropiado, e incluso en los casos en que fuera sencillo (por ejemplo, usando el cuadrado del error de previsión), su distribución asintótica puede ser difícil de derivar. Estos problemas no aparecen si el criterio de penalización de las previsiones es la función discreta. Se volverá a este punto en la introducción de la sección siguiente, cuando se presenten los contrastes que proponemos, y entonces se aclararán estas cuestiones.

3.2.3. Definición formal de la función de pérdida discreta y notación básica

Formalicemos la definición de la función de pérdida discreta, además de presentar otra notación relevante para el desarrollo del documento:

a) Se designará por y_t y v_t el dato y previsión en/para el periodo t y por T el tamaño muestral, siguiendo la misma notación empleada en la Sección 2. El dominio de y y v, llamémosle D, se particiona en las regiones $r_1, r_2, ..., r_m$ ($\bigcup_{i=1}^{k} r_i = D$ y $r_i \cap r_j = \varnothing$, para $i \neq j$) y el conjunto de regiones lo denotamos por R, es decir, $R = \{r_1, r_2, ..., r_m\}$. De este modo, el dominio bidimensional de los datos y previsiones $D \times D$ ha quedado particionado en m^2 "cuadrantes", y cada par (y_t, v_t) tendrá asociado uno de ellos.[10]

b) Designamos por φ a la función $\varphi : D \to R$ que asigna una región de R a cada dato/previsión. Por tanto, los pares (y_t, v_t) tienen asignado por φ un par (r_i, r_j) (cuadrante (i, j)).

[10] En realidad, de cara a los desarrollos de los contrastes posteriores, las particiones efectuadas para los datos y previsiones podrían ser distintas. Sin embargo, tal asimetría no parece muy razonable en la práctica, así que, para favorecer la simplicidad notacional y conceptual del análisis, suponemos que la partición es la misma para ambas variables.

c) Se denotará por n_{ij} la frecuencia del suceso $\{y_t \in r_i, v_t \in r_j\}$, es decir, el número de pares (y_t, v_t) que quedaron situados en el cuadrante dado por (r_i, r_j), mientras se denota por n_i^y y n_j^v las frecuencias de los sucesos $\{y_t \in r_i\}$ y $\{v_t \in r_j\}$, respectivamente. Obviamente, $n_i^y = \sum_{j=1}^{m} n_{ij}$ y $n_j^v = \sum_{i=1}^{m} n_{ij}$.

d) La función de pérdida se define como $g : R \times R \to A$, siendo $A = \{a_1, a_2, ..., a_J\}$, $J \leq m^2$ y $a_1 < a_2 < ... < a_J$. De este modo, dado el par (y_t, v_t), la pérdida resultante será $z_t = g(\varphi(y_t), \varphi(v_t))$. En principio, lo razonable es que la penalización mínima posible tome valor cero, es decir, $a_1 = 0$, y que, por tanto, $a_i > 0$ para $i = 2, ..., J$.

e) Por consiguiente, a partir de la muestra de datos y previsiones, se dispondrá de un conjunto de pérdidas $\{z_1, z_2, ..., z_T\}$, realizaciones de una variable aleatoria Z. A su vez, dicho conjunto puede considerarse como las T realizaciones de un experimento aleatorio con J sucesos posibles, excluyentes entre sí, de probabilidades $p_1, p_2, ..., p_J$. Sea n_i la frecuencia asociada a la pérdida a_i, si suponemos que las realizaciones del experimento son independientes, podríamos afirmar que $(n_1, n_2, ..., n_J)'$ es un vector aleatorio de distribución Multinomial, de parámetros $T, p_1, ..., p_J$.

f) Por último, simbolizaremos por R_s el conjunto de cuadrantes cuya pérdida asignada es a_s, es decir, $R_s = \{(r_i, r_j) | g(r_i, r_j) = a_s\}$.

3.3. Hipótesis nula y alternativa en el contexto planteado

Como ya se ha mencionado, el planteamiento arriba presentado permite diseñar hipótesis que definan más razonablemente el concepto de utilidad respecto a un conjunto de previsiones. En el nuevo contexto basado en pérdidas, lo que querremos contrastar es la posición en la recta real de la distribución asociada a la muestra aleatoria simple de pérdidas $z_1, z_2, ..., z_T$, y dicha posición puede contrastarse a través de la esperanza matemática de la distribución, cuya expresión será $E(Z) = ap$, siendo $a = (a_1, a_2, ..., a_J)$, $p = (p_1, p_2, ..., p_J)'$.[11] Ahora la pregunta es ¿qué valor tendría la media de la distribución de pérdidas si las previsiones no fueran útiles? La respuesta es ambigua, porque lo es la hipótesis "las previsiones no son útiles", pero parece admisible decir que será la misma que la que tendría si los datos y previsiones fueran estocásticamente independientes. Si q denota el valor del vector p bajo independencia estocástica, dicha media será aq. La expresión de q viene dada por (4). La hipótesis alternativa será unilateral: $ap < aq$. De este modo, los contrastes no rechazarán la no utilidad de las previsiones si éstas funcionan *al menos tan mal como* si fueran independientes de los datos, dicho con más precisión, si *generan pérdidas no suficientemente menores que las que generarían unas previsiones independientes de los datos*. Así, no se contrasta independencia estocástica, sino que ésta sirve para diseñar unas pérdidas de referencia (grandes), respecto a las que comparar las pérdidas observadas en la muestra. Además, a diferencia de lo que sucede con TC $m \times m$ y P-T, el uso de una función de pérdidas nos permite definir una hipótesis alternativa unilateral, de modo que el rechazo de la nula implique que las previsiones son razonablemente útiles, en vez de, simplemente, no independientes de los datos.

Los tres primeros contrastes que presentamos a continuación utilizarán las hipótesis citadas, cuya formalización es:

$$H_0 \equiv E(Z) = ap = aq, \tag{2}$$

$$H_1 \equiv E(Z) = ap < aq, \tag{3}$$

$$q_s = \sum_{(r_i, r_j) \in R_s} p_i^y p_j^v, \ s = 1, ..., J, \tag{4}$$

siendo $a = (a_1, a_2, ..., a_J)$, $p = (p_1, p_2, ..., p_J)'$, mientras R_s quedó definido en el punto f) del apartado anterior.

[11] Sería concebible contrastar la posición de la distribución de Z a través, por ejemplo, de la mediana, o a través de otros momentos poblacionales diferentes al primero.

La expresión del vector q se deduce de la suposición de independencia estocástica entre datos y previsiones, teniendo en cuenta la forma de la función de pérdida g definida. Dado que $p_s = \sum_{(r_i,r_j)\in R_s} P(y_t \in r_i, v_t \in r_j)$, entonces $q_s = \sum_{(r_i,r_j)\in R_s} p_i^y p_j^v$, siendo p_i^y y p_j^v las probabilidades marginales $P(y_t \in r_i)$ y $P(v_t \in r_j)$, respectivamente.

El último de los tests que propondremos será diferente a los anteriores. No será un contraste no paramétrico sobre la posición de la distribución de Z en la recta real, sino un contraste paramétrico sobre el vector p de la distribución Multinomial de las frecuencias $(n_1, n_2, ..., n_J)'$, definidas al final del apartado 3.2. La hipótesis nula será independencia entre datos y previsiones, pero la alternativa está definida de modo que implique no solo dependencia estocástica entre datos y previsiones, sino también utilidad de éstas últimas, en el sentido que se expone a continuación:

$$H_0' \equiv p = q, \tag{5a}$$

$$H_1' \equiv H_2 - H_0', \text{ siendo } H_2 \equiv \sum_{j=1}^{i} p_j \geq \sum_{j=1}^{i} q_j, \ i = 1, ..., J, \tag{5b}$$

donde la expresión $H_2 - H_0'$ significa que $H_1^{'}$ está definida por todos los casos que verifiquen H_2, excluyendo aquel que verifica también H_0'.

La idea de fondo de H_2 es que, *agregadamente* (pero no necesariamente cada una), las pérdidas pequeñas sean más probables de aparecer que si datos y previsiones fueran independientes (análogamente, que las pérdidas grandes sean menos probables que bajo independencia, de manera agregada). Para ver esto, recuérdese que p_j es la probabilidad asociada a la j-ésima *menor* pérdida ($a_1 < a_2 < ... < a_J$).

La hipótesis H_1' es más restrictiva que H_1, es decir, $H_1' \rightarrow H_1$ pero no viceversa. Por tanto, para considerar útiles las previsiones, éstas no solo deben generar pérdidas menores *en media* que bajo independencia con los datos, sino que se exige que las probabilidades de las pérdidas generadas mantengan la ordenación descrita en H_2. La implicación $H_1' \rightarrow H_1$ se demuestra en el Apéndice B. Por tanto, podrían tenerse vectores de probabilidades p y q tales que no verifiquen H_1' y, en cambio, sí cumplan H_1, es decir, $ap < aq$. Por ejemplo:

Supongamos que se está haciendo predicción de cierta variable aleatoria, y se utiliza para evaluar las previsiones la función de pérdidas discreta (1), de modo que existen cuatro pérdidas distintas, de valores 0, 1, 2 y 3, respectivamente. Supóngase que las probabilidades poblacionales de las pérdidas son $p = (p_1, ..., p_J)' = (0,30, 0,05, 0,50, 0,15)'$ (es decir, $P(z_t = 0) = 0,30$, $P(z_t = 1) = 0,05$, etc), mientras las teóricas bajo independencia entre datos y previsiones son $q = (q_1, ..., q_J)' = (0,10, 0,40, 0,38, 0,12)'$.[12] Véase que H_1 se cumple: $ap = 1{,}50 < aq = 1{,}52$. Sin embargo, se tiene que $p_1 + p_2 = 0,35 < 0,50 = q_1 + q_2$ y $p_1 + p_2 + p_3 = 0,85 < 0,88 = q_1 + q_2 + q_3$, por lo que no se verifica H_1'. La pérdida media obtenida por el modelo de previsión es menor que si previsiones y datos fueran independientes, pero no es cierto que la suma de las probabilidades de las pérdidas más pequeñas sea mayor que bajo independencia: por ejemplo, la probabilidad de obtener una pérdida menor o igual que 1 es menor que si los datos fueran independientes de las previsiones, y lo mismo ocurre para pérdidas menores o iguales que 2.

[12] Un ejemplo completamente bien diseñado hubiera consistido en idear una matriz 4×4 de probabilidades en los 16 cuadrantes, de donde obtener p y q (q se generaría por (4)). En el caso que mostramos, esto no se ha hecho así, por lo cual esos vectores presentados en el texto pueden no ser coherentes, en el sentido de que posiblemente no exista ninguna matriz de probabilidades 4×4 tal que, valoradas según (1), devuelvan los vectores p y q propuestos. No obstante, para lo que se pretende ilustrar, dicha coherencia es irrelevante.

4. Contrastes propuestos

Como se ha explicado ya en el apartado anterior, lo natural es emplear tests sobre la posición de la distribución de pérdidas en la recta real, como métodos estadísticos para contrastar la utilidad de las previsiones. Las opciones que tenemos son esencialmente dos:

a) La primera sería usar un test no paramétrico estándar de entre los existentes para el contraste de posición de una distribución (Signos y Wilcoxon), que tienen validez en muestras finitas. Dado que el test de Signos resultaría ser poco potente, lo razonable sería emplear el test de Wilcoxon de rangos signados. Sin embargo, está derivado bajo el supuesto de que la distribución es continua, que no es el caso en nuestro contexto.

b) La segunda opción es utilizar un test de significación con validez asintótica, para contrastar $E(Z) = aq$. El Teorema Central del Límite o, análogamente, las propiedades del estimador de Máxima Verosimilitud (MV) para el vector paramétrico p de en una distribución Multinomial, permiten derivar un contraste de este tipo fácilmente. Ofreceremos dos versiones de dicho test, C1-v1 y C1-v2. Además, intentaremos también construir numéricamente la distribución exacta del estadístico de contraste empleado en C1-v1, de modo que dispongamos de una versión del contraste válida en muestras finitas, que denotaremos por C2. Aunque no lograremos caracterizar la distribución exacta, estamos interesados en analizar si la obtenida se aproxima a la verdadera en muestras cortas más que la distribución derivada en los tests asintóticos.

Finalmente, y tal y como anticipamos en el apartado anterior, una alternativa es, en vez de contrastar la posición de la distribución de Z, contrastar el valor del vector paramétrico p en la distribución Multinomial de las frecuencias $(n_1, n_2, ..., n_J)'$, utilizando las hipótesis (5a) y (5b). El contraste será un tipo especial de los tests de Razón de Verosimilitudes, en su versión asintótica.

Antes de pasar a la presentación de los contrastes, recuperemos lo dicho en el punto c) del apartado 3.2.2, cuando explicábamos ventajas de la función de pérdidas discreta. En teoría, podríamos haber diseñado igualmente tests sobre la posición de la distribución de probabilidad de las pérdidas utilizando una función de pérdidas continua. Denotemos una función de pérdidas genérica por $h(y, v)$. La hipótesis nula sería $E(h) = E(h_{IE})$, siendo h_{IE} la función de pérdidas h bajo independencia estocástica. Lo idóneo sería utilizar $\overline{h} - \overline{h}_{IE}$ como estadístico de contraste. Sin embargo:

a) ¿Cómo calculamos $\overline{h}_{IE}$? Obviamente, no hay una muestra que contenga realizaciones de h_{IE}, sino que se necesita conocer la expresión que toma $E(h_{IE})$, para poder calcular $\overline{h}_{IE}$ en base a las realizaciones que se disponen de las variables y y v. Por ejemplo, si utilizamos como función de pérdida el error de previsión al cuadrado $h(y_t, v_t) = (y_t - v_t)^2$, se tiene que $E(h_{IE}) = E[(y_t - v_t)^2] = E(y_t^2) + E(v_t^2) - 2E(y_t)E(v_t)$. En este caso, por tanto, la expresión de $\overline{h}_{IE}$ es perfectamente conocida (combinación de los momentos muestrales asociados a la expresión anterior) y el cálculo puede hacerse sin problemas.

Sin embargo, no podremos obtener una expresión de $E(h_{IE})$ para cualquier función continua h,[13] y funciones como el error de previsión al cuadrado carecen de la riqueza y flexibilidad respecto a la valoración de las previsiones que sí tiene la función discreta (léanse puntos a) y b) del apartado 3.2.2).

b) Incluso aunque se conozca la expresión de $E(h_{IE})$ y se supere el problema descrito, la distribución asintótica del estadístico $\overline{h} - \overline{h}_{IE}$ no será, en general, fácilmente derivable. En este caso, se podría recurrir a utilizar la distribución asintótica de $\overline{h}$ bajo la hipótesis nula, sustituyendo $E(h)$ por $\overline{h}_{IE}$. No es un enfoque correcto en términos teóricos, pero puede ser válido en la práctica. Nosotros lo utilizaremos en C1-v1, para así tener una versión del contraste que pueda implementarse a través de un cálculo muy simple.

En cambio, si h es la función discreta, los problemas a) y b) desaparecen: la expresión de $E(h_{IE})$ es conocida siempre (independientemente de la distribución de la variable bidimensional (y, v)) y, además, el uso de teoremas sobre frecuencias muestrales permitirá obtener fácilmente la distribución asintótica de $\overline{h} - \overline{h}_{IE}$, como se verá en C1-v2.

A continuación se exponen los contrastes propuestos. Se empleará la notación y conceptos introducidos en las subsecciones 3.2 y 3.3. Además de tal notación, se designará por $\widehat{p}$ el estimador MV de p, es decir, $\widehat{p} = (\frac{n_1}{T}, \frac{n_2}{T}, ..., \frac{n_J}{T})'$.

[13] Para conocer la expresión de $E[h(y_t, v_t)]$ bajo independencia entre y_t y v_t, se requiere que la forma funcional de h sea $h(y_t, v_t) = \sum f_k(y_t) g_k(v_t)$.

4.1. Contraste C1-v1

Empezaremos proponiendo un contraste muy sencillo, que sigue el enfoque ya introducido en el punto b) anterior (usar la distribución asintótica de $\overline{h}$ en vez de la de $\overline{h} - \overline{h}_{IE}$). En adelante, volveremos a la notación utilizada para las pérdidas discretas. El contraste es el que sigue:

Como es lógico, para contrastar el valor de la media poblacional de Z utilizaremos su media muestral, $\overline{Z} = T^{-1}a(n_1, n_2, ..., n_J)' = a\widehat{p}$. La derivación del test es simple:

Las hipótesis nula y alternativa del contraste serán H_0 y H_1, especificadas en (2) y (3), respectivamente. Por los Teoremas 1 y 2 citados en el apartado 2.5, sobre la distribución límite del estimador MV en una distribución Multinomial, se tiene que $\sqrt{T}a(\widehat{p} - p) \xrightarrow{L} N(0, aV_pa')$, siendo $V_p = \Omega - pp'$ y Ω una matriz diagonal $J \times J$ con los elementos de p en la diagonal. Utilizando su análogo muestral $\widehat{V}_p = V_p|_{p=\widehat{p}}$ para estimar V_p, el Teorema 4 sobre convergencia en probabilidad y convergencia límite del apartado 2.5 garantiza el siguiente resultado:

$$\sqrt{T}\widehat{W}_p^{-1/2}a(\widehat{p} - q) \xrightarrow[H_0]{L} N(0, 1),$$

siendo $\widehat{W}_p = a\widehat{V}_pa'$.

El problema es que se conoce la expresión del vector q, presentada en (4), pero no su valor numérico. Procederemos sustituyendo q por su estimador consistente $\widehat{q}$, a saber:

$$\widehat{q}_s = \sum_{(r_i,r_j)\in R_s} \widehat{p}_i^y \widehat{p}_j^v, \ s = 1, ..., J, \tag{6}$$

siendo $\widehat{p}_i^y = \frac{n_i^y}{T}$ y $\widehat{p}_j^v = \frac{n_i^v}{T}$ las frecuencias relativas de los datos para la región r_i de la partición y de las previsiones para la región r_j, respectivamente (es decir, los estimadores MV de las probabilidades marginales $P(y_t \in r_i)$ y $P(v_t \in r_j)$). La consistencia de $\widehat{q}$ está garantizada por los puntos (ii)-(iii) del Teorema 4 del apartado 2.5.

Aplicando el punto (i) de dicho Teorema, se deriva finalmente el test a continuación para el contraste de H_0 frente a H_1:

$$C_{1,1} = \sqrt{T}\widehat{W}_p^{-1/2}a(\widehat{p} - \widehat{q}) \xrightarrow[H_0]{L} N(0, 1). \tag{7}$$

El contraste es unilateral, con la cola inferior de la distribución $N(0, 1)$ como región crítica. En concreto, sea $C_{1,1}^0$ el valor observado del estadístico $C_{1,1}$, se rechazará H_0 si $C_{1,1}^0 < \lambda_\alpha$, para un nivel de significación α prefijado, siendo λ_α el punto para el que $\Phi(\lambda_\alpha) = \alpha$, donde Φ denota la función de distribución $N(0, 1)$. Por lo tanto, se rechaza la hipótesis de no utilidad de las previsiones cuando la media muestral de las pérdidas, $\overline{Z} = a\widehat{p}$, es "suficientemente" menor que la pérdida media teórica correspondiente al supuesto de independencia entre datos y previsiones. Como puede verse, el rechazo de la hipótesis nula en este contraste significa no solo que las previsiones no son independientes estocásticamente de los datos sino que son más útiles para el usuario que previsiones independientes de los datos, quedando el concepto de utilidad definido por el usuario a través de la función de pérdida.

4.2. Contraste C1-v2

El problema que surgía en la derivación del test anterior debido al desconocimiento del valor numérico de q puede evitarse enfocando la derivación del contraste para que el estadístico cuya distribución asintótica se deduce sea $a\widehat{p} - a\widehat{q}$ en vez de $\overline{Z} = a\widehat{p}$ (es decir, se trata de utilizar la distribución de $\overline{h} - \overline{h}_{IE}$). De este modo, la media del estadístico de contraste según H_0 ya no es aq, como antes, sino $ap - aq$, con lo que la hipótesis nula nos proporciona precisamente su valor numérico, 0, y ya no es preciso estimar q. Como veremos, se obtendrá un contraste igual que el anterior salvo por la varianza W_p, que será distinta (y ahora denotaremos por G_p). El proceso de derivación del estadístico de contraste y de su distribución asintótica es el siguiente:

Sea p_{ij} la probabilidad de que dato y previsión se sitúen simultáneamente en las regiones r_i, r_j y p_i^y y p_j^v las probabilidades marginales de que el dato se sitúe en la región r_i y la previsión en la región r_j,

respectivamente. Sea $P = (p_{11}, p_{12}, ..., p_{1m}, p_{21}, ..., p_{2m}, ..., p_{m1}, ..., p_{mm})'$ y $\widehat{P}$ su estimador MV (frecuencias relativas):

a) Utilizando de nuevo los Teoremas 1 y 2 del apartado 2.5 sobre la distribución límite del estimador MV de una distribución Multinomial, se tiene que $\sqrt{T}(\widehat{P} - P) \xrightarrow{L} N(0, V_P)$, siendo $V_P = \Omega - PP'$ y Ω matriz diagonal $m^2 \times m^2$ con los elementos de P en la diagonal.

b) El Teorema 3 de dicho apartado sobre convergencias en distribución para funciones diferenciables f permite afirmar que $\sqrt{T}(f(\widehat{P}) - f(P)) \xrightarrow{L} N(0, \nabla f(P) V_P \nabla f(P)')$. Pues bien, utilizando la función f de R^{m^2} en R^J definida por

$$f(P) = p - q = \left(\sum_{(r_i,r_j)\in R_1} p_{ij} - (p_i^y p_j^v), \sum_{(r_i,r_j)\in R_2} p_{ij} - (p_i^y p_j^v), ..., \sum_{(r_i,r_j)\in R_J} p_{ij} - (p_i^y p_j^v) \right)',$$

se tiene que $\sqrt{T}((\widehat{p} - \widehat{q}) - (p - q)) \xrightarrow{L} N(0, \nabla f(P) V_P \nabla f(P)')$, siendo $\nabla f(P)$ la matriz gradiente de la función vectorial f, de dimensiones $J \times m^2$. [14]

c) Ahora puede obtenerse la distribución límite asociada al estadístico que nos interesa, $(a\widehat{p} - a\widehat{q})$: $\sqrt{T}(a(\widehat{p} - \widehat{q}) - a(p - q)) \xrightarrow{L} N(0, G_p)$, siendo $G_p = a \nabla f(P) V_P \nabla f(P)' a'$.

d) El problema que surgía en C1-v1 queda ahora eliminado, ya que el estadístico de contraste es $(a\widehat{p} - a\widehat{q})$, para el que, precisamente, la hipótesis nula proporciona el valor numérico de su esperanza matemática, 0. De este modo, se tiene que $\sqrt{T}(a(\widehat{p} - \widehat{q}))G_p^{-1/2} \underset{H_0}{\xrightarrow{L}} N(0, 1)$.

e) Finalmente, utilizamos el estimador consistente $\widehat{G}_p$ en vez de G_p y, debido al Teorema 4 del apartado 2.5, se mantiene la distribución límite. Por tanto, el test propuesto para el contraste de H_0 frente a H_1 es:

$$C_{1,2} = \sqrt{T}\widehat{G}_p^{-1/2} a(\widehat{p} - \widehat{q}) \underset{H_0}{\xrightarrow{L}} N(0, 1), \tag{8}$$

siendo $\widehat{G}_p = a \left[\nabla f(P)\right]_{P=\widehat{P}} \widehat{V}_P \left[\nabla f(P)\right]'_{P=\widehat{P}} a'$ y $\widehat{V}_P = V_P|_{P=\widehat{P}}$. El valor de los elementos de la matriz $\nabla f(P)$ se presenta en el Apéndice A.

Por supuesto, el contraste es unilateral, y la región crítica se corresponde con la cola inferior de la distribución $N(0, 1)$, como en C1-v1.

La única diferencia entre los contrastes (7) y (8) es la varianza del estadístico de contraste $a\widehat{p} - a\widehat{q}$. Empíricamente hemos comprobado que la desigualdad $\widehat{G}_p < \widehat{W}_p$ se verifica prácticamente siempre (es decir, la estimación de $p - q$ se realiza con más precisión que la estimación de p) y, por tanto, $|C_{1,2}| > |C_{1,1}|$ y el test (8) tendrá una probabilidad de rechazar H_0 en favor de la utilidad de las previsiones mayor que la del test (7). La verdadera varianza asintótica de $\sqrt{T}(a\widehat{p} - a\widehat{q})$ es G_p, no W_p. Sin embargo, el cálculo de W_p es mucho más sencillo que el de G_p, por lo que mantenemos las dos versiones del contraste hasta comprobar en los ejercicios de simulación de la sección 5 qué diferencias se observan entre las propiedades de ambos.

La derivación de este contraste presenta mucha semejanza con la del test P-T. La diferencia entre ambos reside en la introducción de la función de pérdidas en C1-v2, la no equivalencia de sus hipótesis, y en que la comparación entre frecuencias observadas y frecuencias estimadas bajo el supuesto de independencia entre datos y previsiones involucra en C1-v2 a todos los cuadrantes, algo que no sucede en P-T. De todos modos, en cierto caso particular de la función de pérdida discreta, existe una relación explícita entre ambos tests. Esto ocurre cuando la función de pérdida utilizada en C1-v2 consta solo de dos valores: uno (a_1) para la diagonal principal (cuadrantes (i, i)) y otro (a_2) para el resto de cuadrantes. Podemos designar dichas funciones de pérdida por el término "simétricas". En dicha situación, se puede demostrar que las hipótesis nulas de los dos tests son la misma y que sus estadísticos de contraste verifican $C_{1,2} = -S_{PT}$. Tales resultados se muestran en el Apéndice B. Dado que la distribución de contraste es la $N(0, 1)$ en ambos, pero que el test P-T es bilateral mientras C1-v2 es unilateral (siendo la cola de rechazo la inferior), esto significa que, incluso en este caso de máxima similitud entre los dos tests, sus decisiones diferirán muchas veces: P-T rechazará la hipótesis nula si $S_{PT} < \lambda_{\alpha/2}$ ó si $S_{PT} > \lambda_{1-\alpha/2}$, siendo $\Phi(\lambda_{\alpha/2}) = \alpha/2$ y $\Phi(\lambda_{1-\alpha/2}) = 1 - \alpha/2$, donde Φ denota la función de distribución $N(0, 1)$, mientras C1-v2 lo hará solo si $S_{PT} > \lambda_{1-\alpha}$. Es decir, si $S_{PT} \in (-\infty, \lambda_{\alpha/2}) \cup (\lambda_{1-\alpha}, \lambda_{1-\alpha/2})$, las decisiones de estos dos tests serán

[14] Revísese la definición de los conjuntos $R_1, ..., R_J$ realizada en el apartado 3.2.3.

opuestas. Pese a la coincidencia de los estadísticos de contraste (en valor absoluto), de las distribuciones de contraste y de las hipótesis nulas, incluso en este caso, existen asimetrías en las decisiones de C1-v2 y P-T, propiciadas por la diferencia de sus hipótesis alternativas. Por lo tanto, ni siquiera puede decirse que C1-v2 es una generalización de P-T para incorporar funciones de pérdida, o, equivalentemente, que P-T es un caso particular de C1-v2 para funciones de pérdida "simétrica", afirmación que sí hubiera podido hacerse si la hipótesis alternativa de P-T fuera unilateral, siendo la región crítica la cola superior de la distribución.

4.3. Contraste C2

El contraste C1-v1 utiliza la distribución asintótica que corresponde al estadístico $\overline{Z}$ bajo la hipótesis nula H_0. Sin embargo, se podría diseñar un test utilizando la distribución exacta, en vez de la asintótica, con las ventajas que esto pueda suponer cuando las muestras sean cortas. La expresión analítica de la función de masa de dicho estadístico no se conoce, pero sí podemos construir las probabilidades numéricamente. Para ello, nos basaremos en el hecho ya mencionado de que el el vector de frecuencias $(n_1, n_2, ..., n_J)'$ asociadas a las penalizaciones $a_1, a_2, ..., a_J$ $(a_1 < a_2 < ... < a_J)$ definidas por la función de pérdida, sigue una distribución Multinomial de parámetros $T, p_1, p_2, ..., p_J$ (véase apartado 3.2.3). En breve, expondremos el procedimiento para el cálculo de la función de distribución mencionada, que denotaremos por $F_{\overline{Z}}$. El test C2 que proponemos contrastará la hipótesis nula H_0 contra la alternativa H_1 de forma obvia: una vez calculada $F_{\overline{Z}}$, se rechazará la hipótesis nula en favor de la alternativa si y solo si el valor observado del estadístico $\overline{Z} = a\widehat{p}$ es menor que el valor crítico λ_α para un nivel de significación prefijado α, siendo λ_α el valor que verifique $F_{\overline{Z}}(\lambda_\alpha) = \alpha$.[15]

Como veremos a continuación, el cálculo de $F_{\overline{Z}}$ requiere el valor numérico del vector de probabilidades $p = (p_1, p_2, ..., p_J)'$ de la distribución Multinomial asociada a las frecuencias, que habrá de estimarse. Por ello, en realidad, no obtendremos la verdadera distribución $F_{\overline{Z}}$, sino una estimación de ella. En consecuencia, nuestro contraste C2 no será exacto en muestras finitas, pero la distribución utilizada $\widehat{F}_{\overline{Z}}$ sí debería constituir una mejor aproximación a la exacta que la asintótica, en longitudes muestrales pequeñas.

Denotemos el valor observado del estadístico $\overline{Z} = a\widehat{p}$ por $\overline{Z}_0$. Análogamente, sea $Z_T = T\overline{Z} = a(n_1, n_2, ..., n_J)'$ y su valor observado, Z_T^0. El procedimiento de estimación de $F_{\overline{Z}}$ es el que sigue:

(a) Se obtiene el conjunto de todos los posibles vectores de frecuencia asociados al tamaño muestral T. Dicho de manera formal, se trata de obtener el conjunto $\Psi = \{(n_1, n_2, ..., n_J)' | n_1 + n_2 + ... + n_J = T\}$, siendo n_i entero no negativo. Se ha diseñado un algoritmo apropiado para obtener todos estos vectores.

(b) Cada uno de los vectores de frecuencia en el conjunto Ψ se premultiplica por el vector a, obteniéndose el vector D_z, que contendrá todos los posibles valores x_i que puede tomar el estadístico Z_T (cada uno de los cuales aparecerá repetido en D_z cierto número de veces, ya que, en general, habrá varios vectores de frecuencia cuyo producto por el vector a devuelvan el mismo valor x). Sea $\{x_1, x_2, ..., x_L\}$ el conjunto de los valores distintos que aparecen en D_z, es decir, el soporte de la variable aleatoria Z_T.

(c) Para calcular el valor de la función de masa en un punto x_i, $P(Z_T = x_i)$, deben seleccionarse los vectores de frecuencias del conjunto Ψ cuyo valor asociado en D_z fue x_i, es decir, aquellos que verifican $a(n_1, n_2, ..., n_J)' = x_i$, y calcular la probabilidad de que la muestra de longitud T generara cualquiera de los vectores de frecuencia seleccionados. Dicha probabilidad viene determinada por la distribución Multinomial de parámetros $p_1, p_2, ..., p_J, T$. Formalmente:

– Se selecciona de Ψ el subconjunto $\Theta = \{(n_1, n_2, ..., n_J)' \in \Psi | Z_T = a(n_1, n_2, ..., n_J)' = x_i\}$.

– $P(Z_T = x_i) = P((n_1, n_2, ..., n_J)' \in \Theta) = \sum_{(n_1, n_2, ..., n_J) \in \Theta} P_M(n_1, n_2, ..., n_J | p_1, ..., p_J, T)$, siendo $P_M(.)$ la función de masa de una distribución Multinomial, a saber:

$$P_M(n_1, n_2, ..., n_J | p_1, ..., p_J, T) = \prod_{i=1}^{J} (p_i)^{n_i} \frac{T!}{n_1! n_2! ... n_J!}. \tag{9}$$

Como ya se mencionó arriba, para el cálculo de las probabilidades en (9), se requiere una estimación del vector p. Dado que lo que se pretende es obtener la distribución del vector de frecuencias bajo la hipótesis

[15] En realidad, no habrá ningún valor que verifique tal igualdad (salvo por casualidad). Por ello, en teoría, debería aplicarse una versión aleatorizada del test. Este asunto se trata en el Apéndice A. No obstante, para la exposición que aquí sigue, este detalle puede ignorarse.

nula H_0, no procede emplear $\widehat{p}$, sino un vector que verifique dicha hipótesis. Lógicamente, utilizaremos la estimación del vector q (véase la expresión (6)), vector que cumple la restricción exigida por H_0.

(d) Se define la estimación de la función de distribución asociada a la variable aleatoria Z_T por: $\widehat{F}_{Z_T}(x_i) = \sum_{s=1}^{i} P(Z_T = x_s)$, para los puntos del soporte $\{x_1, x_2, ..., x_L\}$.

(e) Finalmente, la variable aleatoria $\overline{Z} = Z_T/T$ tendrá como función de distribución $F_{\overline{Z}}(y_i) = F_{Z_T}(x_i)$ para los puntos de su suporte, que será el conjunto $\{y_1, y_2, ..., y_L\} = \{x_1/T, x_2/T, ..., x_L/T\}$. Con todo el proceso, por tanto, habremos logrado la estimación deseada $\widehat{F}_{\overline{Z}}$.

Dicha distribución dependería no solo de los parámetros $p = (p_1, p_2, ..., p_J)'$ y T, sino también del vector a. Por lo tanto, debe construirse explícitamente para cada vector $(p', a', T)'$. En realidad, no deben calcularse todos los puntos de la distribución, sino solo los necesarios para la realización del contraste, es decir, hasta que se haya identificado el valor crítico λ_α. Dicho formalmente, solo se estimarán los puntos $F_{\overline{Z}}(y_1)$, $F_{\overline{Z}}(y_2)$, ..., $F_{\overline{Z}}(\lambda_\alpha)$, siendo $F_{\overline{Z}}(\lambda_\alpha) = \alpha$.

El contraste expuesto presenta la importante ventaja de ser (casi) válido para muestras finitas, mientras su problema radica en la dificultad de cálculo, en términos de costo computacional. Es fácilmente demostrable que la cardinalidad del conjunto Ψ es $CR(J,T) = C(T+J-1,T) = \frac{T+J-1!}{T!J-1!} = \frac{(T+J-1)(T+J-2)...(T+1)}{J-1!}$, donde CR y C designan combinaciones con y sin repetición, respectivamente. De acuerdo a esto, cualquier algoritmo para el cálculo de los puntos necesarios de $F_{\overline{Z}}$ (pasos (a)-(e)) presenta un coste computacional de orden de magnitud T^{J-1}, lo que puede convertir el test en poco operativo para valores de T ó J relativamente elevados.

Además, en principio, el contraste no podría llevarse a cabo si alguna componente del vector $\widehat{q}$ resultara igual a cero (porque el valor de la expresión sería siempre nulo, o, dicho de otro modo, porque, de acuerdo a los resultados muestrales, la dimensión definida para el vector de frecuencias es excesiva). En la práctica, podemos evitar este obstáculo procediendo de forma sencilla tal y como se explica en el apartado 4.6.

4.4. Contraste C3

Como ya hemos venido anticipando en el apartado 3.3 y en la introducción de éste, nuestra última propuesta consiste en diseñar un contraste paramétrico sobre el valor del vector p, aprovechando que se conoce la función de probabilidad de las frecuencias $(n_1, ..., n_J)$. Lo adecuado es emplear un test de Razón de Verosimilitudes, en versión asintótica. La primera idea que viene a la cabeza es utilizarlo del modo estándar, es decir, con hipótesis alternativa igual a la negación de la hipótesis nula. La especificación formal de ésta sería $p = q$. Sin embargo, la especificación de la hipótesis alternativa conduciría a problemas semejantes a los que ya hemos apuntado para TC $m \times m$ y P-T en el apartado (3.1). Un test así rechazaría la hipótesis de no utilidad de las previsiones también cuando éstas estuvieran correladas negativamente con los datos. Lo que procede entonces es un test RV especial, cuya hipótesis alternativa no sea la simple negación de la hipótesis nula.

Precisamente, Robertson y Wright (1981) derivan dos tests para contrastar hipótesis sobre ordenación estocástica de probabilidades de una población Multinomial. El que nos interesa es el primero de ellos. Éste utiliza una hipótesis nula simple contra una hipótesis alternativa según la cual las probabilidades de la población Multinomial presentan una determinada ordenación. El estadístico utilizado es el de Razón de Verosimilitudes y Robertson y Wright obtienen la expresión del estimador MV bajo la hipótesis alternativa y la distribución asintótica del estadístico RV, que no será ya Chi-Cuadrado. El contraste encaja perfectamente en el contexto que tratamos, definido en la subsección 3.2, y de su aplicación en dicho contexto resulta el test que proponemos a continuación:

La hipótesis nula será $H_0' \equiv p = q$. Pero la hipótesis alternativa es ahora H_1', cuya formulación es (5b). Tal y como se comentó en el apartado 3.3, ésta expresa una condición de utilidad para las previsiones más estricta o exigente que la definida para los tests anteriores, H_1. Consúltese el apartado citado para una interpretación y algunos comentarios sobre dicha hipótesis.

Bajo estas hipótesis, el estadístico Razón de Verosimilitudes y su distribución asintótica son:

$$C_3 = 2\sum_{i=1}^{J} n_i(\ln \overline{p}_i - \ln \widehat{q}_i), \tag{10}$$

donde $\widehat{q}$ es el estimador (6), mientras $\overline{p}$ representa el estimador MV de p bajo H_1'. El estimador $\overline{p}$ en absoluto es estándar, y fue derivado por Robertson y Wright (1981). Más adelante volveremos a ello. Debido a que H_0 es un punto en la frontera de H_1', los resultados habituales Chi-cuadrado no se cumplen y la distribución de C_3 viene definida, para cualquier número real t, por:

$$\lim_{T\to\infty} P(C_3 \geq t) = \sum_{l=1}^{J} P(l,J) P(\chi^2_{J-l} \geq t), \tag{11}$$

donde χ^2_v denota una variable Chi-cuadrado con v grados de libertad ($\chi^2_0 \equiv 0$) y $P(l,J)$ es una probabilidad cuyo significado y cálculo se tratará en breve.

Sea C_3^0 el valor observado de C_3, se rechazará H_0 en favor de H_1' cuando se verifique $P(C_3 \geq C_3^0) < \alpha$, siendo α un nivel de significación prefijado, y habiéndose calculado la probabilidad mediante (11). Fijémonos que el resultado del contraste no depende en ningún modo de los valores numéricos de las pérdidas $a_1, ..., a_J$, a diferencia de lo que ocurre para los dos tests propuestos anteriormente. Por supuesto, sí depende de la estructura de la función de pérdidas, es decir, del número de éstas y de su disposición en los cuadrantes de la partición del dominio bidimensional de datos y previsiones.

La implementación del test presenta cierta complejidad, debido a lo poco convencional de la definición de $\overline{p}$ y a las dificultades de cálculo de $P(l,J)$.

En primer lugar, la coordenada i-ésima de $\overline{p}$ es el producto entre la coordenada i-ésima de $\widehat{p} = (n_1/T, ..., n_J/T)$ y la coordenada i-ésima de cierta "proyección" del vector $q/\widehat{p} = (q_1/\widehat{p}_1, ..., q_J/\widehat{p}_J)$ sobre un conjunto $C = \left\{q/\widehat{p} \in R^J | q_1/\widehat{p}_1 \geq q_2/\widehat{p}_2 \geq ... \geq q_J/\widehat{p}_J\right\}$, proyección ésta que utiliza $\widehat{p}$ como vector de "pesos". Es decir:

$$\overline{p}_i = \widehat{p}_i E^{(i)}_{\widehat{p}}(q/\widehat{p}|C),\ i = 1, 2, ..., J, \tag{12}$$

siendo $E^{(i)}_{\widehat{p}}(q/\widehat{p}|C)$ la coordenada i-ésima de la proyección $E_{\widehat{p}}(q/\widehat{p}|C)$. Dicha proyección no representa sino un vector de ponderaciones adecuadas que aplicar sobre cada elemento de $\widehat{p}$, de manera que se garantice que $\overline{p}$ verifica la restricción $H_1^{'}$ formulada en (5b), pero modificando lo menos posible la estimación original MV $\widehat{p}$ (de hecho, si $\widehat{p}$ ya la verificaba, $E_{\widehat{p}}$ resultará un vector de unos, y, por tanto, se tendrá que $\overline{p} = \widehat{p}$). La ponderación $E^{(i)}_{\widehat{p}}$ es siempre uno de los ratios $\frac{\widehat{q}_1}{\widehat{p}_1}$, $\frac{\widehat{q}_1+\widehat{q}_2}{\widehat{p}_1+\widehat{p}_2}$, ..., $\frac{\widehat{q}_1+...+\widehat{q}_i}{\widehat{p}_1+...+\widehat{p}_i}$. No obstante, el cálculo exacto de la proyección se obtiene aplicando el "Lower Sets Algorithm" (LSA), que utilizan Robertson y Wright (1981). La descripción del algoritmo LSA, así como algunos ejemplos para ilustrar su funcionamiento en la generación de la proyección $E_{\widehat{p}}$, se exponen con detalle en el Apéndice A. La demostración de que (12) es la estimación MV bajo H_1' se ofrece en Robertson y Wright (1981), en su Teorema 2.1. Finalmente, debe añadirse que, de acuerdo a dicho teorema, la expresión en (12) es correcta si $\widehat{p}_i > 0$, para $i = 1, 2, ..., J$, es decir, si $n_i > 0$. De incumplirse esta condición, nuestra propuesta es obrar de forma simple en el sentido que se expondrá en el apartado 4.6, igual que comentamos anteriormente para los tests TC 2×2, TC $m \times m$ y C2, cuando presentaban un problema equivalente.

Por otro lado, las probabilidades $P(l,J)$ que forman parte del cálculo en (11) son las probabilidades de que $E_{\widehat{p}}(q/\widehat{p}|C)$ tenga exactamente l valores distintos (ver detalles del Algoritmo LSA) y la forma de obtener estos valores se expone también en el Apéndice A, utilizando los resultados ofrecidos por Barlow et al. (1972). El cómputo de $P(l,J)$ se simplifica considerablemente si $q = (J^{-1}, ..., J^{-1})$, para $J \leq 12$. De no ser así, estas probabilidades son complicadas de calcular pero se obtienen a través de una relación recursiva (ecuación (3.23) de Barlow (1972)) si $J \leq 4$. Para otros casos, $P(C_3 \geq t)$ puede aproximarse utilizando los límites obtenidos por Robertson y Wright (1980):

$$\lim_{T\to\infty} P(C_3 \geq t) \leq \frac{1}{2}\left(P(\chi^2_{J-1} \geq t) + P(\chi^2_{J-2} \geq t)\right). \tag{13}$$

El empleo del límite superior (13) daría lugar a un test más conservador (en en sentido de que la probabilidad de rechazar la hipótesis nula sería menor) que el que se tiene si se calculan las probabilidades exactas (11).[16]

[16] En la implementación de C3 que hacemos para nuestras simulaciones se utiliza la distribución asintótica correcta (11).

Las diferencias conceptuales entre este contraste y el tipo de contraste anterior (en cualquiera de sus tres versiones, C1-v1, C1-v2 ó C2) radican, esencialmente, en la especificación de las hipótesis, especialmente la alternativa, más exigente ahora para valorar como útiles las previsiones, y en la independencia de C3 respecto de los valores numéricos concretos de la función de pérdidas. Ésta última característica puede ser favorable o desfavorable para el usuario según la aplicación (aumenta la robustez del test pero disminuye la flexibilidad de la definición de utilidad que introduce el usuario a través de la función de pérdidas). Por su parte, la peculiar hipótesis alternativa tenderá a disminuir la probabilidad de rechazo de la hipótesis de ausencia de utilidad de las previsiones, respecto a los otros tests.

4.5. Ejemplos de aplicación de los contrastes propuestos

El Cuadro 1 presenta los resultados de los contrastes que proponemos y de los dos tests utilizados en la literatura para contrastar utilidad de un conjunto de previsiones en el contexto de cambio direccional en particiones de $m \geq 2$ regiones, para los tres conjuntos de previsiones expuestos en la sección 3.1. El Cuadro 1 muestra los valores obtenidos para los estadísticos de contraste, los valores que toma la función de distribución correspondiente en dichos estadísticos y la decisión del contraste en cada caso, rechazo (R) o no rechazo (NR) de la hipótesis nula de "previsiones no útiles". Para interpretar correctamente los resultados del Cuadro 1, recuérdese que la región crítica de los tests TC y C3 es la cola superior, la inferior en el caso de C1-v1, C1-v2 y C2, y ambas en el caso del contraste P-T.

Si repasamos las matrices de frecuencias detalladas en la sección 3.1, y tal y como se mencionó entonces, es obvio que las previsiones generadas por el modelo M1 carecen de utilidad para un usuario, las del modelo M3 son excelentes y las de M2 probablemente serán útiles, aunque dependerá de la función de pérdida del individuo y la penalización que asigne a aciertos en el signo del cambio de la variable pero que tengan asociados errores en magnitud.

Para llevar a cabo los contrastes, se ha supuesto una función de pérdida cuyas asignaciones se corresponden con la matriz (1) presentada en el apartado 3.2, función que penaliza una previsión certera en signo con una pérdida máxima de 1, y una incorrecta en signo con un valor mínimo de 2.

Los resultados obtenidos son favorables a nuestros contrastes. Mientras los tests usuales, TC y P-T, rechazan la hipótesis nula en todos los casos y los valores de los estadísticos de contraste no se muestran sensibles a las obvias diferencias entre los conjuntos de previsiones asociados a M1, M2 y M3 (salvo P-T en el caso de M3), los tests C1-v1, C1-v2, C2 y C3 se comportan como uno esperaría de un contraste sobre utilidad de previsiones en un contexto de cambio de dirección de una variable (para el que la valoración de previsiones implícita en la matriz (1) sea adecuada): no rechazan la hipótesis de no utilidad de las previsiones del modelo M1 y rechazan dicha hipótesis en los casos de los modelos M2 y M3. Los valores de los estadísticos en el Cuadro 1 muestran que no existe duda en la decisión de ninguno de los contrastes.

Cuadro 1. Resultados de los contrastes para los ejemplos de la sección 3.1

	Valor observado del estadístico de contraste					
	TC 4×4 (S_{TCm})	P-T (S_{PT})	C1-v1 ($C_{1,1}$)	C1-v2 ($C_{1,2}$)	C2 ($\overline{Z}= a\widehat{p}$)	C3 (C_3)
M1	292,31	-353,55	19,90	40,62	2,50	0,00
M2	292,31	-353,55	-49,25	-17,60	1,00	208,31
M3	292,31	74,94	-74,12	-43,77	0,00	265,61
	Valor Función Distribución en el valor observado del estadístico contraste					
	TC 4×4	P-T	C1-v1	C1-v2	C2	C3
M1	1,0	0,0	1,0	1,0	1,0	0,06
M2	1,0	0,0	0,0	0,0	0,0	1,0
M3	1,0	1,0	0,0	0,0	0,0	1,0
	Decisión del contraste					
	TC 4×4	P-T	C1-v1	C1-v2	C2	C3
M1	R	R	NR	NR	NR	NR
M2	R	R	R	R	R	R
M3	R	R	R	R	R	R

Nota: R denota rechazo de la hipótesis nula, y NR, no rechazo

Antes de seguir adelante, vamos a mostrar otro ejemplo que permita comprender el papel determinante que juega en el resultado de nuestros contrastes el concepto de utilidad de previsiones que maneje el usuario, en su aplicación concreta, concepto que se introducirá en el test a través de la definición de la función de pérdida. Supongamos que estamos en un nuevo contexto predictivo en el que la precisión es fundamental. Pensemos, por ejemplo, en un hospital con cierto tipo de enfermos, con cuatro posibles enfermedades. Todas las enfermedades tienen síntomas similares, pero las enfermedades A y B son del mismo tipo pero variante diferente, y lo mismo ocurre entre C y D. Se debe diagnosticar correctamente la enfermedad y poner el tratamiento adecuado. Una equivocación supone haber aplicado un tratamiento incorrecto, agravando la salud del enfermo por pérdida de tiempo hasta rectificar, por lo que tiene un coste muy alto siempre. La rectificación de tratamiento es más compleja cuanto más diferentes fueran los tratamientos aplicados respecto a los correctos, lo que depende (no de forma lineal) de con qué otra enfermedad se confundió. Se tienen tres mecanismos de diagnóstico (M1, M2, M3), cuyos resultados se quieren evaluar sobre una muestra de 100 pacientes. La siguiente función de pérdidas puede reflejar la percepción de los médicos sobre la utilidad de los métodos de diagnóstico:[17]

		v_t			
		A	B	C	D
y_t	A	0	1,75	2	3
	B	1,75	0	2	3
	C	3	2	0	1,75
	D	3	2	1,75	0

(14)

Es decir, se trata de la misma función de pérdida que en el ejemplo anterior, solo que los cuadrantes que se correspondían entonces con aciertos en el signo del dato aunque no precisos en magnitud (ahora significa confundir una enfermedad con la otra variante del mismo tipo), pasan de ser penalizados con una pérdida numérica 1, a serlo con pérdida 1.75. Fijémonos que la penalización sobre confusiones con la primera de las enfermedades de tipo diferente es poco mayor que confundir con la enfermedad del mismo tipo pero diferente variante. Podríamos justificar esto de muchos modos. En cualquier caso, se hace así para facilitar la comparación con la función de pérdida del ejemplo anterior, por lo que no alargaremos la exposición con justificaciones técnicas. Supongamos que los resultados de los tres métodos de diagnóstico vuelven a poder resumirse a través de las matrices M_1, M_2 y M_3. Pues bien, el Cuadro 2 muestra la actuación de los contrastes ahora, con las mismas muestras de datos y previsiones que antes pero con la nueva función de pérdida (14). Por supuesto, los tests TC y P-T son invariantes a este cambio, ya que no utilizan función de pérdida. Los nuestros mantienen los mismos resultados que en el caso anterior para los métodos M1 y M3, porque el cambio en las pérdidas solo afecta a situaciones de confusión entre enfermedades del mismo tipo, situaciones éstas que no ocurrieron en ninguna ocasión con estos dos modelos (se produce algún cambio en el valor de los estadísticos de contraste de C1-v1 y C1-v2 porque la modificación del valor numérico de la pérdida en cuestión sí afecta a la estimación de las varianzas involucradas en los estadísticos ($\widehat{W}_p$ y $\widehat{G}_p$, respectivamente)). Sin embargo, respecto a M2, las decisiones de los contrastes C1-v1, C1-v2 y C2 se han invertido en relación al ejemplo anterior. Los tres contrastes pasan ahora a no rechazar la hipótesis nula,[18] considerando, por tanto, las previsiones de M2 no útiles, porque no son suficientemente precisas (aunque los estadísticos de contraste muestran que los tests, implícitamente, identifican diferencias favorables a M2 respecto a M1). En cambio, los resultados de C3 se mantienen inalterados respecto al ejemplo anterior, ya que, como se comentó cuando se presentó el test en el apartado 4.4, éste es sensible a la estructura de la función de pérdidas (número de pérdidas y su disposición en los cuadrantes), pero no al valor numérico de las mismas.

[17] En este caso, y_t se refiere a la enfermedad verdadera del paciente t-ésimo, y v_t a la diagnosticada por el hospital.

[18] Suponiendo que el nivel de significación del contraste es menor que 0,72.

Cuadro 2. Resultados de los contrastes para el ejemplo de la sección 4.5

	Valor observado del estadístico de contraste					
	TC 4×4 (S_{TCm})	P-T (S_{PT})	C1-v1 ($C_{1,1}$)	C1-v2 ($C_{1,2}$)	C2 ($\overline{Z}= a\widehat{p}$)	C3 (C_3)
M1	292,31	-353,55	16,15	33,46	2,50	0,00
M2	292,31	-353,55	26,13	2,56	1,75	208,31
M3	292,31	74,94	-83,54	-48,95	0,00	265,61
	Valor Función Distribución en el valor observado del estadístico contraste					
	TC 4×4	P-T	C1-v1	C1-v2	C2	C3
M1	1,0	0,0	1,0	1,0	1,0	0,06
M2	1,0	0,0	1,0	0,99	0,72	1,0
M3	1,0	1,0	0,0	0,0	0,0	1,0
	Decisión del contraste					
	TC 4×4	P-T	C1-v1	C1-v2	C2	C3
M1	R	R	NR	NR	NR	NR
M2	R	R	NR	NR	NR	R
M3	R	R	R	R	R	R

Nota: R denota rechazo de la hipótesis nula, y NR, no rechazo.

Por último, presentamos en el Cuadro 3 las pérdidas mínimas a_2 que serían necesarias en los cuadrantes (1,2), (2,1), (3,4) y (4,3) (los correspondientes a previsiones certeras en signo pero no precisas en magnitud −primer ejemplo−, o diagnósticos correctos en el tipo de enfermedad pero no en su variante concreta −segundo ejemplo−) para no rechazar la hipótesis de no utilidad de las previsiones generadas por un modelo como M2 (es decir, con resultados iguales a los de M_2), en cada uno de nuestros tests salvo C3, según nivel de significación aplicado:

Cuadro 3.
Valor mínimo de la pérdida $(1, 2)$
en la función (1) para no rechazar
H_0 con las frecuencias M_2

α	C1-v1	C1-v2	C2
0,20	1,66	1,64	1,54
0,10	1,66	1,62	1,48
0,05	1,66	1,61	1,42
0,01	1,66	1,59	1,32

4.6. Algunos detalles de implementación de los tests

En la práctica, la implementación de algunos de los contrastes presentados en las dos secciones anteriores requiere ciertos detalles que no especificamos entonces para no hacer farragosa la exposición con cuestiones que no son centrales para el entendimiento de los tests. Desarrollaremos ahora estas cuestiones para que el usuario de los contrastes pudiera implementarlos de forma completa. El lector poco interesado en aplicar los tests en la práctica puede pasar por alto este apartado.

I. En primer lugar, abordaremos el asunto de las *frecuencias nulas*. Como se ha ido mencionando al exponer los tests, TC 2×2, TC $m \times m$, C2 y C3 no podrían ser implementados si, dada la muestra utilizada, alguna frecuencia relativa muestral, de las necesarias en el cálculo de los estadísticos de contraste, resultara nula (repásense los apartados correspondientes a estos cuatro tests para saber a qué sucesos se corresponden dichas frecuencias en cada test). Para evitar tener que renunciar a la aplicación del contraste, nuestra propuesta es muy sencilla: sustituir la verdadera frecuencia relativa muestral nula por un valor muy pequeño, llamémosle ξ,[19] reconstruyendo convenientemente el resto de elementos del vector de frecuencias relativas de modo que su suma siga siendo uno. En concreto, si denotamos por r el vector de dimensión l de frecuencias relativas, el mecanismo exacto que estamos empleando es el siguiente: sustituir cada elemento

[19] Aconsejamos valores del tipo $\xi = 0,001$.

$r_i = 0$ por $r'_i = \xi$ y cada $r_j > 0$ por $r'_j = r_j - \frac{s_1 \xi}{l - s_1}$, siendo s_1 el número de elementos de r que resultaron nulos en el vector.

Creemos que éste es el modo de lograr aplicar el test distorsionando lo menos posible la información de las observaciones muestrales.

II. En segundo lugar, debe tratarse la problemática de la *imposibilidad de construir regiones críticas de tamaño exactamente igual a* α y la consiguiente posible solución, la *aleatorización*. Este asunto afecta solamente a tests cuyo estadístico de contraste es una variable aleatoria discreta. En nuestro caso, por tanto, concierne a los tests HM, B y C2.

La descripción del problema y de la solución teórica (construcción de tests aleatorizados) son bien conocidos en estadística, y se presentan en el Apéndice A. En nuestras simulaciones, los tres tests citados se implementan usando el procedimiento de aleatorización, tal y como es recomendable para este tipo de experimentos. En cambio, en la aplicación práctica (es decir, cuando se usa el test en muestras reales, no en ejercicios de simulación), los contrastes de distribución discreta no suelen ser implementados bajo su versión aleatorizada, sino que, en los casos de ambigüedad en la decisión del contraste −que son los que resuelve la aleatorización−, suele ser preferirse que el usuario tome la decisión, redefiniendo el riesgo de error de tipo I que desea asumir. En el Apéndice A se detallan también estas cuestiones. No obstante, el problema de los tests con distribución discreta es más irrelevante conforme la distribución es más parecida a una continua, situación que ocurre en los contrastes HM y B al aumentar la longitud muestral, mientras en C2 sucede a mayor tamaño muestral y también a mayor número de pérdidas establecidas en la función de pérdida discreta (es decir, conforme mayor es J).

III. Finalmente, haremos mención al problema de la *independencia de las observaciones muestrales*. Los contrastes que hemos ido proponiendo a lo largo de la sección anterior están suponiendo, a la hora de derivar las distribuciones para el estadístico de contraste, que las T pérdidas generadas a partir de la muestra de datos y previsiones (ie, el conjunto $\{z_1, ..., z_T\}$, según la notación del apartado 3.2.3) son mutuamente independientes. Lo mismo se está suponiendo para ciertas variables relacionadas directamente con la muestra de datos y previsiones (en concreto, aquellas a partir de las que se construye el estadístico de contraste) en el caso de los tests habituales de la literatura presentados en la sección 2.[20] Si no se verificara el supuesto de independencia, los contrastes deberían aplicarse sobre submuestras de observaciones independientes modificando el nivel de significación de cada test individual según la cota de Bonferroni, procedimiento que detallamos en el siguiente capítulo de esta tesis. Ya que a lo largo de este capítulo usaremos solamente casos en donde la condición de independencia sea asumible,[21] no introducimos aquí dicho método, con la intención de no complicar innecesariamente la exposición.

[20] Por ejemplo, en el test Binomial, es obvio que son las variables $d_1, ..., d_n$ las que deben verificar independencia.

[21] En realidad, de los cuatro experimentos de simulación que se realizan en el capítulo, dos de ellos verifican la condición de independencia en las pérdidas por construcción, pero no así los otros dos. Sin embargo, en el Apéndice C se presentan resultados que evidencian que dicha condición también es asumible en la práctica en esos experimentos, de modo que no es necesario aplicar el procedimiento de Bonferroni.

5. Ejercicios de Simulación

Una vez presentados todos los contrastes anteriores, vamos a comprobar cómo se comportan en distintos escenarios predictivos y a comparar dichos comportamientos. Para ello, generaremos datos según algún PGD, preveremos a través de un modelo de previsión estimado − que puede o no tener la misma especificación que el PGD − y estimaremos la probabilidad de rechazar la hipótesis nula de "no utilidad de las previsiones", que está asociada implícitamente a cada test. El análisis que se realizará no será un estudio estándar de tamaño y potencia de contrastes, y esto es, en última instancia, debido a la ambigüedad de la hipótesis nula citada. Es decir, en cada ejercicio de simulación, no estará claro si la hipótesis es cierta o falsa, y, por tanto, no hay una idea a priori definida y unánime sobre qué debería hacer el test, si rechazar o no rechazar, sino que dependerá del juicio del usuario y de lo que considere "razonable". Por ejemplo, si el PGD es un MA(1) y se usa para predecir un AR(1), ¿las previsiones son útiles (hipótesis nula falsa) o no (hipótesis nula cierta)? ¿debería el test haber rechazado la hipótesis de partida o no? La respuesta no es evidente. Si acaso, podemos evaluar si los tests verifican o no propiedades "razonables" del siguiente tipo: cuando la correlación entre previsiones y datos sea positiva y elevada, es deseable que el test presente una probabilidad de rechazo lo mayor posible, y lo contrario si la correlación es positiva pero muy débil, o negativa. Nuevamente, existirá una ambigüedad en la cuantificación numérica de los adjetivos "elevada" o "débil", pero éste es, inevitablemente, el ámbito en el que nos tendremos que manejar.

Para entender las diferencias en la forma de proceder entre los tests de la literatura y los propuestos por nosotros, además de comparar las probabilidades de rechazo de la hipótesis nula en los experimentos de simulación, vamos a aprovechar éstos para identificar casos en los que la decisión de todos los contrastes de la literatura (rechazo o no de la hipótesis nula) sea justo la opuesta a la de todos nuestros tests. Ésta será una información muy interesante, y que, desde nuestro punto de vista, es esclarecedora sobre la ventaja de usar los contrastes que proponemos.

Debe recordarse que los contrastes utilizan diferentes definiciones formales de la hipótesis conceptual de "no utilidad de las previsiones", es decir, interpretan dicha hipótesis de distinta manera: por ejemplo, el test Binomial considera previsiones no útiles si la probabilidad de acierto en el signo no es superior a 0,5; los tests TC, si datos y previsiones son independientes estocásticamente; P-T utiliza una definición similar a la anterior aunque algo menos restrictiva, etc; finalmente, los contrastes propuestos en este documento la identifican con igualdad entre la esperanza matemática de las pérdidas y la que se obtendría si datos y previsiones fueran independientes.[22] Por ello, las diferencias entre las propiedades de todos estos tests provendrán fundamentalmente de las diferencias en la definición de su hipótesis nula, más que de otras razones técnicas. En este sentido, tendemos a pensar que nuestra definición es más acertada que la de los tests tradicionales y nuestro objetivo ahora es esencialmente obtener evidencia a favor de que dicha definición da lugar a comportamientos más "razonables" en nuestros tests que los que tienen los tests habitualmente empleados en la literatura sobre evaluación de previsiones.

5.1. Diseño de los experimentos

5.1.1. Definición de los PGD y modelos de previsión

Se van a llevar a cabo cuatro experimentos de simulación:

Experimento 1 (MA(1) vs MA(1)):

PGD: $y_t \sim$ MA(1): $y_t = \varepsilon_t - \theta\varepsilon_{t-1}$, $|\theta| < 1$, $\varepsilon_t \overset{iid}{\sim} N(0, \sigma_\varepsilon^2)$.

MODELO DE PREVISIÓN: $v_t \sim$ MA(1): $v_t = -\widehat{\theta}\widehat{\varepsilon}_{t-1}$, siendo $\widehat{\theta}$ la estimación del parámetro en un modelo MA(1) y $\widehat{\varepsilon}_{t-1} = y_{t-1} - v_{t-1}$.

Como ya se ha hecho notar arriba, el coeficiente de correlación lineal contemporánea entre datos y previsiones, ρ_{yv}, puede ser una medida de referencia de la bondad de las últimas, igual que podría serlo el ratio entre el error cuadrático medio de las previsiones y la varianza de los datos, $\frac{ECM_{vy}}{\sigma_y^2}$. Debido a la linealidad de todos los modelos empleados, en los cuatro experimentos que llevamos a cabo se verifica la relación $\frac{ECM_{vy}}{\sigma_y^2} = 1 - \rho_{yv}^2$, de modo que las dos medidas son equivalentes.

[22] Esto no es exacto en el caso del test C3. Véase (5a) en el apartado 3.3.

Bajo la aproximación $\widehat{\theta} \approx \theta$ (aproximación perfectamente asumible en nuestro diseño porque las estimaciones se realizarán siempre con muestras de longitud muy elevada, tal y como se precisará en el apartado 5.1.2), es inmediato comprobar que en el Experimento 1 el coeficiente ρ_{yv} tiene la expresión $\rho_{yv} = \varphi(\theta) = \left| \frac{\theta}{\sqrt{1+\theta^2}} \right|$. La función φ está definida para $\theta \in (-1, 1)$ y toma valores en $(0, \frac{1}{\sqrt{2}})$, es simétrica respecto de $\theta = 0$, y creciente para $\theta \in (0, 1)$. De modo que su mínimo es $\varphi(0) = 0$ y su máximo, $\varphi(1) = \frac{1}{\sqrt{2}} \approx 0,7$.

Realizaremos el experimento bajo valores diferentes valores positivos de θ y, por tanto, de ρ_{yv}, siendo esperable que la probabilidad de rechazo de los tests aumente al aumentar θ. Los valores elegidos son $\theta = 0,204$, $0,314$, $0,436$, $0,577$, $0,750$ y $0,927$, generando datos y previsiones correlados por $\rho_{yv} = 0,20$, $0,30$, $0,40$, $0,50$, $0,60$ y $0,68$, respectivamente.

Experimento 2 (MA(1) vs AR(1)):

PGD: $y_t \sim$ MA(1): $y_t = \varepsilon_t - \theta\varepsilon_{t-1}$, $|\theta| < 1$, $\varepsilon_t \overset{iid}{\sim} N(0, \sigma_\varepsilon^2)$ y $\sigma_\varepsilon^2 = 0,1$.

MODELO DE PREVISIÓN: $v_t \sim$ AR(1): $v_t = \widehat{\phi} y_{t-1}$, siendo $\widehat{\phi}$ la estimación del parámetro en un modelo AR(1).

En este caso, la expresión de ρ_{yv} es $\rho_{yv} = \varphi(\theta) = \frac{|\theta|}{1+\theta^2}$, donde, además de la aproximación $\widehat{\phi} \approx \phi$, se ha tenido en cuenta que ϕ coincide con el coeficiente de autocorrelación simple de orden 1 de y_t y, en un modelo MA(1), éste es $\phi = -\frac{\theta}{1+\theta^2}$. La función φ está definida para $\theta \in (-1, 1)$ y toma valores en $(0, \frac{1}{2})$, es simétrica, y creciente en $(0, 1)$. Por tanto, su mínimo es $\varphi(0) = 0$ y su máximo, $\varphi(1) = \frac{1}{2}$.

Ahora, los valores elegidos son $\theta = 0,209$, $0,333$, $0,500$ y $0,905$, generando datos y previsiones correlados por $\rho_{yv} = 0,20$, $0,30$, $0,40$ y $0,49$, respectivamente.

Experimento 3 (Mod. lineal 2 variables explicativas vs Mod. lineal 1 variable explicativa):

PGD: $y_t = \beta_1 x_{1t} + \beta_2 x_{2t} + \varepsilon_t$, $w_t = (x_{1t}, x_{2t}, \varepsilon_t)' \overset{iid}{\sim} N(0_{3\times 1}, \Sigma)$, siendo $\Sigma = \begin{pmatrix} \sigma_1^2 & 0 & 0 \\ 0 & \sigma_2^2 & 0 \\ 0 & 0 & \sigma_\varepsilon^2 \end{pmatrix}$. Fijaremos, para todas las ejecuciones de este experimento, los siguientes valores paramétricos: $\beta_1 = \beta_2 = \sigma_1^2 = 1$ y $\sigma_\varepsilon^2 = 0,1$.

MODELO DE PREVISIÓN: $v_t = \widehat{\beta}_1 x_{1t}$, siendo $\widehat{\beta}_1$ la estimación del parámetro del modelo lineal, mal especificado, ya que incluye solo x_{1t} como variable explicativa.

La expresión de ρ_{yv} queda $\rho_{yv} = \varphi(\sigma_2^2) = \frac{|\beta_1\sigma_1|}{\left(\beta_1^2\sigma_1^2 + \beta_2^2\sigma_2^2 + \sigma_\varepsilon^2\right)^{1/2}} = \frac{1}{\sqrt{1,1+\sigma_2^2}}$ (se ha usado la aproximación $\widehat{\beta}_i \approx \beta_i$). La función φ está definida para $\sigma_2^2 \in (0, \infty)$ y toma valores en $(0, \frac{1}{\sqrt{1,1}})$, y es decreciente en todo su dominio. Así que su mínimo es $\varphi(\infty) = 0$ y su máximo, $\varphi(0) = \frac{1}{\sqrt{1,1}} \approx 0,95$.

Hemos realizado el experimento para los siguientes valores: $\sigma_2^2 = 23,9$, $10,01$, $2,90$, $0,94$, $0,134$ y $0,008$, generando datos y previsiones correlados por $\rho_{yv} = 0,20$, $0,30$, $0,50$, $0,70$, $0,90$ y $0,95$, respectivamente.

Experimento 4 (Normal Bivariante):

Finalmente, llevaremos a cabo un experimento que no requerirá realizar estimación. En este caso, los datos y las previsiones constituyen un vector bivariante de distribución Normal, con coeficiente de correlación ρ. Es decir:

$(y_t, v_t) \sim N(0_{2\times 1}, \Sigma)$, siendo $\Sigma = \begin{pmatrix} \sigma_y^2 & \rho\sigma_y\sigma_v \\ \rho\sigma_y\sigma_v & \sigma_v^2 \end{pmatrix}$. Fijaremos en todos los ejercicios $\sigma_y^2 = \sigma_v^2 = 1$.

Obviamente, ahora $\rho_{yv} = \rho$.

Este experimento presenta las siguientes ventajas:

– Por propiedades bien conocidas de la distribución Normal bidimensional, podemos identificar independencia estocástica con $\rho = 0$. De este modo, fijando en el experimento $\rho = 0$, puede estimarse el tamaño de los contrastes.[23]

– Permite generar cómodamente datos y previsiones correlados negativamente (simplemente fijando un valor $\rho < 0$). Éste es un caso interesante, puesto que la definición de la hipótesis nula de los contrastes

[23] Porque, aunque solo algunos tests definen la hipótesis nula a través de la independencia estocástica entre datos y previsiones, las hipótesis nulas especificadas por el resto de contrastes se verifican si existe dicha independencia.

TC y PT posiblemente conduzca a rechazarla en favor de la hipótesis de que las previsiones son útiles, algo muy poco razonable. En cambio, esperamos que nuestros tests actúen correctamente, nuevamente gracias a la especificación de sus hipótesis.

Así, realizaremos el experimento bajo los siguientes diseños alternativos: $\rho_{yv} = \rho = 0,0,\ 0,2,\ 0,3,\ 0,4,\ 0,5,\ 0,7,\ 0,9,\ 0,95,\ -0,25$ y $-0,8$.

El cuadro a continuación sintetiza los diseños fijados para los cuatro experimentos a realizar:

Cuadro 4. Resumen Diseños Ejercicios Monte Carlo. Experimentos 1 a 4

	Experimento 1		Experimento 2		Experimento 3		Experimento 4	
	θ	$\rho_{yv} = \left\lvert \frac{\theta}{\sqrt{1+\theta^2}} \right\rvert$	θ	$\rho_{yv} = \left\lvert \frac{\theta}{1+\theta^2} \right\rvert$	σ_2^2	$\rho_{yv} = \frac{1}{\sqrt{1,1+\sigma_2^2}}$	ρ	$\rho_{yv} = \rho$
Diseño 1	0,204	0,20	0,209	0,20	23,90	0,20	0,00	0,00
Diseño 2	0,314	0,30	0,333	0,30	10,01	0,30	0,20	0,20
Diseño 3	0,436	0,40	0,500	0,40	2,90	0,50	0,30	0,30
Diseño 4	0,577	0,50	0,905	0,49	0,94	0,70	0,40	0,40
Diseño 5	0,750	0,60			0,134	0,90	0,50	0,50
Diseño 6	0,927	0,68			0,008	0,95	0,70	0,70
Diseño 7							0,90	0,90
Diseño 8							0,95	0,95
Diseño 9							-0,25	-0,25
Diseño 10							-0,80	-0,80

5.1.2. Resto del diseño

Cada uno de los experimentos descritos se lleva a cabo para muestras de previsiones de cuatro longitudes, $T = 10$, 25, 50 y 100. Los modelos de previsión se estiman cada periodo, en base a toda la información pasada, con un número mínimo de datos R. Es decir, se genera una muestra inicial de R datos con el PGD. En cada periodo $t = R + k$, se utilizan los datos $\{y_s\}_{s=1}^{t}$ para estimar el modelo de previsión y construir la predicción para y_{t+h}, operación que se repite desde $k = 1$ a $k = T - R$. En todos los casos, se empleará $R = 1000$ y horizonte de previsión uno ($h = 1$). El uso de un valor de R tan elevado garantiza que la aproximaciones $\widehat{\theta} \approx \theta$, $\widehat{\phi} \approx \phi$ y $\widehat{\beta}_i \approx \beta_i$ realizadas arriba sean correctas y, sobre todo, sirve para eludir el problema de la incertidumbre paramétrica en tests para evaluar o comparar previsiones. La incertidumbre paramétrica afectaría a las propiedades estadísticas de los contrastes de este capítulo, y será un tema abordado en el Capítulo 3 de la Tesis.

Los modelos de regresión del Experimento 3 y el modelo AR(1) del Experimento 2 se estiman usando MCO. El modelo MA(1) del Experimento 1 se estima utilizando un procedimiento analítico (que puede consultarse, por ejemplo, en Novales (1993), capítulo 13), reduciendo sustancialmente los costes computacionales respecto a cualquier otro método numérico, y sin perjuicio en los resultados de previsión respecto a dichos métodos, dado el elevado número de observaciones muestrales para la estimación.[24] En el caso de los modelos de regresión, las variables exógenas x_{1t} y x_{2t} se consideran deterministas, así que su valor x_{1t+h}, x_{2t+h} es conocido en t por el modelo de previsión.

Se llevan a cabo 1000 repeticiones de cada ejercicio. Los nueve contrastes introducidos hasta ahora se ejecutan para cada una de las repeticiones, cada uno de los tamaños muestrales y cada uno de los diseños presentados, tomando en todos los casos $\alpha = 0,05$ como nivel de significación.

Respecto a las cuestiones de implementación de los tests expuestas en el apartado 4.6, éstos se aplican siguiendo las técnicas allí detalladas. Solo la última de aquellas cuestiones requiere ahora una explicación especial. Las pérdidas que se generan en los contrastes deberían verificar aleatoriedad para que la derivación de los tests fuera correcta. El tipo de no aleatoriedad que se podría sospechar en estos ejercicios de simulación es el asociado a existencia de autocorrelación en las pérdidas. De existir autocorrelación, los contrastes deberían implementarse sobre submuestras no autocorreladas usando la cota de Bonferroni,

[24] Este procedimiento puede no ser idóneo si $\theta \approx 1$, pero no es el caso en ninguno de nuestros diseños del Experimento 1 (el máximo valor de θ utilizado es $\theta = 0,927$, y éste es un valor suficientemente lejano de 1 como para generar problemas).

método que se explicará en el próximo capítulo de la Tesis. Pues bien, en los Experimentos 3 y 4, datos y previsiones, y, por tanto, también las pérdidas, están libres de autocorrelación, por construcción. Pero no ocurre lo mismo en los Experimentos 1 y 2. Por ejemplo, el coeficiente de autocorrelación de orden uno de un proceso MA(1) es $\gamma_y(1) = -\frac{\theta}{1+\theta^2}$, valor que, en nuestros experimentos, se encontraría en el intervalo $[0,2, 0,5]$. Sin embargo, la autocorrelación de las pérdidas $\gamma_z(1)$ será bastante menor. En las Tablas 6-7 del Apéndice C se presentan las estimaciones promedio de dicho coeficiente obtenidas en las simulaciones. Además, en cada repetición de los experimentos, se lleva a cabo un test contrastando la hipótesis $\gamma_z(1) = 0$.[25] En las Tablas 8-9 del Apéndice C se adjunta el porcentaje de veces que no se rechazó dicha hipótesis. Estos resultados sugieren que la hipótesis de no autocorrelación en las pérdidas era asumible y que, por tanto, la forma en la que se van a implementar los contrastes en los Experimentos 3 y 4, sin procedimiento de Bonferroni, es correcta.

5.1.3. Partición del dominio de los datos y función de pérdida

Todos los contrastes requieren elegir una partición del dominio de los datos/previsiones, es decir, una partición de la recta real. En el caso de los tests de dos regiones, la elección es obvia: $(-\infty, 0)$ y $(0, +\infty)$. Para el resto de contrastes (TC $m \times m$, P-T y los cuatro propuestos en este documento), se ha decidido dividir la recta real en tres regiones, a saber: $r_1 = (-\infty, -b\sigma_y)$, $r_2 = (-b\sigma_y, +b\sigma_y)$, $r_3 = (+b\sigma_y, +\infty)$.[26] Se eligió $m = 3$ por simplicidad y la forma de separar las regiones en base a σ_y nos parece razonable. Se decidió tomar $b = 0,4$ para todos los experimentos. Utilizando dicho valor, la probabilidad marginal de que los datos se sitúen en cada una de las regiones r_1, r_2 y r_3 es siempre $0,35$, $0,30$ y $0,35$,[27] respectivamente, mientras la probabilidad marginal de que las previsiones se sitúen en cada región depende del diseño particular de cada experimento,[28] y se adjuntan en el Cuadro 5, abajo (en realidad, se presentan solo las probabilidades correspondientes a la región central r_2, $P(v_t \in r_2)$. Las correspondientes a r_1 y r_3 serán iguales entre sí, e iguales al valor $\frac{1}{2}(1 - P(v_t \in r_2))$). El cálculo de dichas probabilidades es obvio teniendo en cuenta que las distribuciones marginales de y_t y v_t son, en todos los experimentos, $N(0, \sigma_y^2)$ y $N(0, \sigma_v^2)$, respectivamente, por lo que no nos extendemos en hacerlo explícito.

Finalmente, los contrastes que utilizan una función de pérdidas discreta, es decir, los cuatro propuestos por nosotros, usarán la estructura de pérdidas siguiente:

		v_t		
		r_1	r_2	r_3
y_t	r_1	0	1	2
	r_2	2	0	1
	r_3	2	1	0

(15)

Véase que la función de pérdidas elegida establece una asimetría a la hora de penalizar previsiones cuando los datos correspondieron a la región r_2, ya que, en tal caso, las previsiones grandes negativas se penalizan más que las grandes positivas.

Además de la partición y función de pérdidas recién descritas, también llevaremos a cabo los experimentos bajo una partición de cuatro regiones y una especificación de función de pérdidas natural para dicho caso: la partición será la misma que en el caso de tres regiones, solo que dividiendo el espacio central en dos, según el signo de los datos/previsiones, es decir: $r_1 = (-\infty, -b\sigma_y)$, $r_2 = (-b\sigma_y, 0)$, $r_3 = (0, +b\sigma_y)$, $r_4 = (+b\sigma_y, +\infty)$, manteniéndose el valor $b = 0,4$. La matriz de pérdidas será la siguiente (es la matriz que se utilizó en el ejemplo del apartado 4.5):

[25] La especificación del test se presenta en el Apéndice C.

[26] Greer (2003) utiliza, para una aplicación práctica, una partición en tres regiones similar a la que proponemos aquí, incluyendo valores positivos y negativos en la región central, solo que las regiones se delimitan a partir de valores numéricos fijos, independientemente de la desviación típica de la variable a prever.

[27] Siendo exactos, dichas probabilidades son 0,3446, 0,3108 y 0,3446.

[28] Salvo en el Experimento 4, en el que las probabilidades son independientes del diseño, e iguales a las asociadas a los datos, es decir, 0,35, 0,30 y 0,35.

$$
\begin{array}{cc|cccc}
 & & \multicolumn{4}{c}{v_t} \\
 & & r_1 & r_2 & r_3 & r_4 \\ \hline
 & r_1 & 0 & 1 & 2 & 3 \\
y_t & r_2 & 1 & 0 & 2 & 3 \\
 & r_3 & 3 & 2 & 0 & 1 \\
 & r_4 & 3 & 2 & 1 & 0
\end{array} \tag{16}
$$

Esta especificación sigue una lógica muy razonable para aplicaciones donde la previsión del signo es muy relevante: si las previsiones aciertan el signo del dato, son penalizadas con una pérdida máxima 1 si han errado en cuanto a la magnitud de éste; si equivocan el signo, serán penalizadas con pérdida 2 ó 3, según el grado de error en cuanto a la magnitud del dato.

Las probabilidades de rechazo de la hipótesis nula obtenidas en el caso de partición de cuatro regiones y función de pérdidas (16) solo se muestran en el Apéndice D, ya que no aportan información relevante adicional a la que se desprende del caso de tres regiones. En la introducción a esta sección mencionábamos otro tipo de información muy interesante que se obtendría a través de las simulaciones. Se trata de las realizaciones concretas en que los tests de la literatura deciden de forma opuesta a los propuestos en este documento. Pues bien, esta información se recabará tanto en el caso de tres regiones como en el de cuatro, pero será la obtenida en el de cuatro la que presentaremos en detalle, ya que la función de pérdidas (16) permite realizar interpretaciones más intuitivas.

Cuadro 5. Probabilidades Marginales $P(v_t \in r_2)$. Experimentos 1 a 4

	Experimento 1	Experimento 2	Experimento 3	Experimento 4
Expresiones para $P(v_t \in r_2)$				
σ_y^2	$(1+\theta^2)\sigma_\varepsilon^2$	$(1+\theta^2)\sigma_\varepsilon^2$	$1,1+\sigma_2^2$	1
σ_v^2	$\theta^2\sigma_\varepsilon^2$	$\frac{\theta^2}{(1+\theta^2)^2}\sigma_y^2$	1	1
$P(y_t \in r_2)$	$1-2\Phi(-b)$	$1-2\Phi(-b)$	$1-2\Phi(-b)$	$1-2\Phi(-b)$
$P(v_t \in r_2)$	$1-2\Phi(-s)$ $s=b\frac{\sqrt{1+\theta^2}}{\theta}$	$1-2\Phi(-s)$ $s=b\frac{1+\theta^2}{\theta}$	$1-2\Phi(-s)$ $s=b\sqrt{1,1+\sigma_2^2}$	$1-2\Phi(-b)$
Valores Numéricos de $P(v_t \in r_2)$				
Diseño 1	0,955	0,955	0,955	0,311
Diseño 2	0,818	0,818	0,817	0,311
Diseño 3	0,683	0,683	0,576	0,311
Diseño 4	0,576	0,579	0,432	0,311
Diseño 5	0,495		0,343	0,311
Diseño 6	0,443		0,326	0,311
Diseño 7				0,311
Diseño 8				0,311
Diseño 9				0,311
Diseño 10				0,311

Nota 1: Φ simboliza la Función Distribución (cdf) $N(0,1)$; Nota 2: Consultar diseños en Cuadro 4.

5.2. Resultados

Separamos los resultados de los experimentos de simulación en dos partes, según el tipo de análisis. Por un lado, ofrecemos el estudio de las probabilidades de rechazo de la hipótesis nula estimadas en los cuatro ejercicios, que, como ya se dijo en la introducción de la sección, deben tener una interpretación un tanto especial. Por otro, analizaremos casos concretos, correspondientes a realizaciones de los ejercicios de simulación, en los que la decisión de los contrastes propuestos en el documento (nos restringiremos a C1-v1, C1-v2 y C2) resultó ser opuesta a la decisión que se desprende de los contrastes habituales en esta literatura (nos centraremos solamente en aquellos que admiten más de dos regiones, TC $m \times m$ y P-T), es decir, que todos los primeros rechacen la hipótesis nula cuando ninguno de los segundos lo haga, o viceversa.

5.2.1. Probabilidades de Rechazo

En las Tablas 1-4 se presentan las probabilidades de rechazo de la hipótesis nula estimadas para cada test de los evaluados, en los Experimentos 1 a 4. Se corresponden al caso de partición de 3 regiones y función de pérdida (15). Los resultados para la partición de 4 regiones y función de pérdida (16) son similares, y conducen a las mismas conclusiones que se pueden deducir con la información aquí presentada. Por ello, y para evitar redundancias, las tablas de probabilidades de rechazo correspondientes a dicho caso se adjuntan en el Apéndice D.

Las conclusiones fundamentales que pueden deducirse son las siguientes:

1) En primer lugar, los tests son razonablemente exactos en tamaño, salvo C3, que lo infraestima, y P-T, que lo sobreestima en longitudes muestrales muy pequeñas. Esto puede comprobarse en el primer diseño del Experimento 4, único caso de entre todos los experimentos en el que la hipótesis nula *formal* de todos los tests es cierta, debido a que $\rho_{yv} = 0$.

2) Los valores de probabilidad de rechazo de la hipótesis nula están directamente vinculados al coeficiente de correlación ρ_{yv} entre datos y previsiones. Por ejemplo, fijado un valor de ρ_{yv}, no se observan prácticamente diferencias entre los resultados de los tests en el Experimento 1, donde el modelo de previsión coincide con el PGD, y los del Experimento 2, donde difiere de él. Por ello, las conclusiones del comportamiento de los tests pueden obtenerse simplemente restringiéndonos al estudio del Experimento 4, que, por facilidad de diseño, es el más exhaustivo en cuanto a niveles de correlación analizados.

3) Cuando los datos y las previsiones están correlados negativamente (últimos dos diseños del Experimento 4), los tests TC 2×2, TC 3×3 y P-T rechazan con alta probabilidad la hipótesis nula de ausencia de utilidad de las previsiones, en contra de cualquier lógica admisible. Este comportamiento era esperado, ya que su hipótesis nula *formal* es que los datos y previsiones presentan independencia estocástica y la alternativa es simplemente la negación de la nula. Efectivamente, en estos dos diseños, hay dependencia estocástica entre datos y previsiones, pero de ahí no se deduce que éstas sean útiles. Ya se había sugerido este mal funcionamiento en la motivación y ejemplo expuestos en el apartado 3.1. Precisamente, los contrastes que proponemos en este documento salvan este problema, al añadir una estructura de pérdidas en los cuadrantes generados por la partición (véase que sus probabilidades de rechazo en Tabla 4 para los dos diseños citados son nulas, como es deseable). La hipótesis alternativa de nuestros tests solo es cierta cuando las pérdidas obtenidas son *mejores* que las que se obtendrían bajo independencia estocástica, y no cuando son simplemente *distintas*. Como ya dijimos en la introducción de esta sección, es la especificación de las hipótesis de los tests la que es determinante en este contexto para la bondad de los mismos, más que otras cuestiones técnicas.

4) Cuando la correlación entre datos y previsiones es muy alta, sería deseable que los tests rechazaran la hipótesis nula de inutilidad de las previsiones, ya que se trata de una situación clara de hipótesis nula falsa. Podríamos centrarnos en los resultados de los diseños $\rho_{yv} = 0,9$ y $\rho_{yv} = 0,7$ del Experimento 4 para chequear el funcionamiento de los tests en estos contextos:

– En el caso $\rho_{yv} = 0,9$, todos los tests presentan potencia prácticamente uno para muestras de longitud $T \geq 25$. Pero si $T = 10$, la potencia de los tests TC 3×3 y HM solo alcanza valores de 0,55 y 0,66, respectivamente, mientras la de TC 2×2 y P-T se sitúa en 0,73 y 0,76, respectivamente. De los tests habituales en la literatura, solo el Binomial (B) alcanza una potencia de 0,80. En cambio, los contrastes que proponemos obtienen mejores resultados: C1-v1, C1-v2 y C2 logran una potencia de 0,88, 0,88 y 0,82, respectivamente. Únicamente el test C3 se queda en niveles de potencia similares a los de P-T.

– En el caso $\rho_{yv} = 0,7$ (que también es a priori un escenario para el que lo razonable sería considerar útiles las previsiones), la potencia de los tests se sitúa muy próxima a uno en muestras de longitud $T \geq 50$. Por el contrario, cuando $T = 25$, situación bastante más frecuente en la práctica, la potencia de los contrastes de la literatura se reduce a niveles en torno a 0,70, salvo los tests B y HM, que alcanzan potencias de 0,79 y 0,77, respectivamente. Por su parte, la potencia de los tests propuestos en el documento es próxima a 0,90, exceptuando, de nuevo, C3, cuya probabilidad de rechazo estimada fue 0,75.

5) El análisis se torna menos claro cuando la correlación entre datos y previsiones es intermedia, en el sentido sugerido en la introducción a la sección. ¿A partir de qué valor de ρ_{yv} sería deseable que los tests rechazaran? Es decir, ¿a partir de qué valor de ρ_{yv} se consideran las previsiones útiles? Obviamente, no existe una respuesta general. Dependerá del tipo de problema o contexto en que se encuentre el usuario del contraste. A continuación presentamos un Cuadro (Cuadro 6) con el valor ρ^*_{yv} a partir del cual la

probabilidad de rechazo de un test, estimada usando los resultados del Experimento 4, supera 0,50 y 0,90, respectivamente.[29] Para agilizar la exposición, diremos que un test presenta "tendencia al rechazo" y "rechazo casi seguro" si su probabilidad de rechazar la hipótesis nula supera 0,50 y 0,90, respectivamente.

Los resultados muestran que, salvo el test C3, cuyo comportamiento en situaciones donde ρ_{yv} toma un valor "intermedio" es similar al del test P-T, el resto de nuestros contrastes requieren valores de ρ_{yv} menores que los tests tradicionales para presentar tanto tendencia al rechazo de la hipótesis de previsiones no útiles como rechazo casi seguro.

Analicemos primero la tendencia al rechazo de los tests. La diferencia entre los valores ρ^*_{yv} para nuestros contrastes y los de los contrastes de la literatura disminuye con la longitud muestral. Para $T = 50$ y para $T = 100$, los tests actúan de modo menos dispar. Si agrupamos los resultados de los tests tradicionales salvo el test B, podría decirse que, para $T = 100$, los tests tradicionales comienzan a rechazar por encima del 50 % de las ocasiones si $\rho_{yv} \approx 0,30$, en media, mientras para C1-v1, C1-v2 y C2 bastaría una correlación $\rho_{yv} \approx 0,24$. Para $T = 50$, estos valores podrían situarse en $\rho_{yv} \approx 0,41$ y $\rho_{yv} \approx 0,33$, respectivamente. Pero las diferencias más acusadas se producen en longitudes muestrales más cortas, es decir, precisamente en las situaciones más habituales en la práctica. Los tests tradicionales presentan tendencia al rechazo de la hipótesis de no utilidad de las previsiones si éstas mantienen una correlación mínima con los datos de $0,57$ y $0,80$ en muestras de tamaño $T = 25$ y $T = 10$, respectivamente, mientras los contrastes que proponemos (salvo C3) requieren para ello correlaciones menores, $0,45$ y $0,67$, respectivamente, valores que, en principio, parecen más adecuados. Por su parte, el test B actúa de modo similar a C2.

Obviamente, el nivel de correlación necesario para que los tests actúen con rechazo casi seguro es muy superior al necesario para tendencia al rechazo. Ahora, el comportamiento del test B es más próximo al resto de tests tradicionales que en el caso de tendencia al rechazo, y más alejado de C2 que entonces, por lo que las medias de resultados de tests tradicionales que mencionamos a continuación se realizan incluyendo también dicho test. Bien, en media, dichos contrastes rechazan casi seguro la hipótesis de no utilidad de previsiones solo cuando el coeficiente de correlación lineal supera aproximadamente 0,97, 0,83, 0,65 y 0,47 para longitudes muestrales $T = 10$, 25, 50 y 100, respectivamente. Por su parte, la misma media pero calculada para los tests C1-v1, C1-v2 y C2 toma valores 0,92, 0,74, 0,59, 0,40, de nuevo inferiores que los de los contrastes de la literatura. Ahora, las discrepancias entre ambos tipos de tests son menores que cuando analizábamos su tendencia al rechazo, pero las diferencias casi no se amortiguan con el tamaño muestral.

6) Como se acaba de exponer, no existe un valor de correlación entre datos y previsiones que separe los casos en donde las previsiones deben considerarse no útiles para el usuario y aquellos en los que sí. Por arrojar algo más de luz, exponemos en las Tablas 14 y 15 del Apéndice E la media muestral de las pérdidas que se observó en las simulaciones, bajo nuestra función de pérdidas (15), y la pérdida media que se obtendría bajo la hipótesis de independencia estocástica entre datos y previsiones. Es decir, se trata de estimaciones de las variables ap y aq que aparecen en la definición de la hipótesis (2), y que forman parte de los estadísticos de contraste de los tests C1-v1, C1-v2 y C2. Como puede observarse en dichas tablas, siempre que $\rho_{yv} > 0$, se tiene que $\overline{Z} = a\widehat{p} < a\widehat{q}$. En el mismo sentido que se comentó en el punto anterior, es interesante preguntarse cuánto más pequeña debe ser la media muestral de las pérdidas que la que se obtendría si la hipótesis nula fuera cierta, para que la probabilidad de rechazo de los tests supere 0,50 y 0,90, respectivamente. La respuesta se adjunta en el Cuadro 7, cuya construcción y contenido es del mismo tipo al del Cuadro 6, solo que en vez del coeficiente de correlación "crítico" ρ^*_{yv}, se presenta el ratio $\frac{a\widehat{p}}{a\widehat{q}}$ a partir del cual el porcentaje de rechazos de los tests supera el 50 % y 90 %, respectivamente. Las conclusiones son análogas a las obtenidas en el punto 5. Los tests se comportan de modo similar para muestras grandes ($T \geq 50$), pero en longitudes muestrales cortas los tests tradicionales necesitan discrepancias entre ap y aq considerablemente mayores que los nuestros para rechazar. Por ejemplo, si $T = 10$, los contrastes tradicionales (promediando los resultados de todos ellos) requieren ratios $\frac{a\widehat{p}}{a\widehat{q}} \approx 0,45$ y $0,19$ para presentar tendencia al rechazo de la hipótesis nula y rechazo casi seguro, respectivamente, mientras a los propuestos por nosotros, salvo C3, les bastan ratios bastante mayores, $\frac{a\widehat{p}}{a\widehat{q}} \approx 0,60$ y $0,29$. Cuando $T = 25$, los niveles medios requeridos para el ratio en cuestión por los tests tradicionales se elevan a $0,66$ (tendencia al rechazo) y $0,41$ (rechazo casi seguro). Los de C1-v1, C1-v2 y C2 se sitúan en $0,73$ y $0,51$.

[29] Para calcular ρ^* (fijado T) utilizamos interpolación lineal entre las dos probabilidades de rechazo estimadas con las simulaciones $P_1(\rho_1)$ y $P_2(\rho_2)$ tales que $P_1(\rho_1) < r < P_2(\rho_2)$, siendo $r = 0,5$ ó $r = 0,9$.

Todos estos datos de los Cuadros 6 y 7 solo pretenden ser orientativos respecto a las diferencias entre el comportamiento de los dos conjuntos de contrastes que evaluamos, cuando la correlación no es ni negativa, ni nula, ni positiva muy elevada (escenarios ya comentados en los puntos anteriores 1), 3) y 4)). En el siguiente apartado se aportará información sobre la actuación de los tests en casos concretos, donde se describe de forma más detallada la realización de la muestra de datos y previsiones (en vez de simplificarse al valor teórico de ρ_{yv} o del ratio $\frac{ap}{aq}$), lo que permitirá enjuiciar más claramente cuál debería haber sido la decisión de los contrastes, y, por tanto, evaluar mejor las diferencias entre éstos.

7) Los contrastes TC $m \times m$ y P-T han mostrado sus debilidades en casos en que $\rho_{yv} < 0$. Por su parte, los tests que utilizan una partición de solo dos regiones (B, H-M y TC 2×2) tienen un problema obvio: solo califican las previsiones por su capacidad de prever el signo correctamente, pero la magnitud de los errores de previsión no es relevante, algo que en muchas aplicaciones puede ser no aceptable. Aunque no sería necesaria ninguna prueba de esto, adjuntamos en la Tabla 5 los resultados de un experimento, que hemos denotado por Experimento 5, con el mismo diseño que el Experimento 4, pero fijando $\rho_{yv} = 0,7$ en todos los casos, y, en cambio, manipulando la varianza de las previsiones σ_v^2. Uno esperaría que, dada una varianza de los datos σ_y^2 y una correlación entre datos y previsiones ρ_{yv}, fijas ambas, los tests tendieran a mantener más veces la hipótesis de no utilidad de las previsiones conforme menos varianza tengan éstas. Sin embargo, como puede observarse en la Tabla 5, los contrastes de dos regiones son insensibles a los cambios de varianza de las previsiones, ya que esto no afecta a la probabilidad de prever correctamente el signo de los datos.

8) De los resultados empíricos expuestos hasta este punto pueden deducirse ya bastantes conclusiones sobre la comparación entre los nueve tests considerados para contrastar si un conjunto de previsiones es útil, que resumimos a continuación:

a) Los contrastes de dos regiones (B, H-M, TC 2×2) presentan el problema explicado en el punto 7: solamente evalúan las previsiones por su capacidad de predecir correctamente el signo del dato, sin distinguir en absoluto por la magnitud del error de previsión. En todas aquellas situaciones predictivas en las que la magnitud del error de previsión es relevante, estos tres contrastes son poco potentes.

b) Debido a la especificación de sus hipótesis (fundamentalmente, de la hipótesis alternativa), los contrastes TC $m \times m$ y P-T llevan a considerar válidas previsiones cuya correlación con los datos ρ_{yv} es negativa, algo inadmisible en cualquier aplicación.

c) El test C3 se comporta de modo similar a P-T cuando $\rho_{yv} > 0$, pero con la ventaja de no cometer la incorrección mencionada en b) cuando $\rho_{yv} < 0$.

d) Los otros tres contrastes introducidos en el documento, C1-v1, C1-v2 y C2, presentan propiedades parecidas en la mayor parte de las situaciones. En muestras muy cortas pueden apreciarse diferencias, siendo C1-v2 el que presenta normalmente mayor probabilidad de rechazo de la hipótesis nula y C2, el que menos de los tres. C2 tiene la desventaja de ser más costoso computacionalmente.

Cuadro 6. Correlación ρ^* a partir de la que Prob Rechazo > 0,50, 0,90

T	B	H-M	TC 2×2	TC $m \times m$	P-T	C1-v1	C1-v2	C2	C3
Primer caso: Probabilidad Rechazo > 0,50									
10	0,71	0,78	0,77	0,87	0,76	0,66	0,63	0,72	0,77
25	0,48	0,51	0,56	0,61	0,59	0,45	0,44	0,47	0,55
50	0,36	0,37	0,41	0,44	0,44	0,33	0,32	0,34	0,39
100	0,26	0,26	0,30	0,32	0,33	0,24	0,24	0,24	0,30
Segundo caso: Probabilidad Rechazo > 0,90									
10	0,95	0,99	0,99	0,98	0,95	0,91	0,91	0,93	0,95
25	0,81	0,82	0,84	0,84	0,84	0,74	0,72	0,76	0,83
50	0,62	0,63	0,66	0,66	0,68	0,59	0,57	0,60	0,65
100	0,45	0,44	0,48	0,48	0,51	0,40	0,40	0,41	0,47

ρ^* se calcula con resultados de Exp. 4, interpolando entre las probs más próximas a 0,50, 0,90

Cuadro 7. Nivel ratio $\frac{a\widehat{p}}{a\widehat{q}_0}$ a partir del que Prob Rechazo > 0,50, 0,90

T	B	H-M	TC 2×2	TC $m\times m$	P-T	C1-v1	C1-v2	C2	C3
Primer caso: Probabilidad Rechazo > 0,50									
10	0,54	0,46	0,48	0,36	0,49	0,59	0,61	0,53	0,48
25	0,72	0,70	0,66	0,62	0,64	0,74	0,74	0,72	0,66
50	0,79	0,79	0,76	0,74	0,74	0,81	0,81	0,80	0,77
100	0,85	0,86	0,83	0,82	0,82	0,87	0,87	0,86	0,83
Segundo caso: Probabilidad Rechazo > 0,90									
10	0,23	0,15	0,16	0,16	0,23	0,31	0,31	0,26	0,23
25	0,44	0,42	0,40	0,39	0,39	0,52	0,54	0,49	0,41
50	0,61	0,60	0,58	0,58	0,57	0,64	0,65	0,63	0,59
100	0,74	0,74	0,72	0,72	0,70	0,77	0,77	0,76	0,73

$\frac{a\widehat{p}}{a\widehat{q}_0}$ se calcula con resultados de Exp. 4, interpolando entre las probs más próximas a 0,50, 0,90

Tabla 1. Probabilidad Rechazo Experimento 1 (PGD: MA(1); Modelo Estimado: MA(1))
$y_t = \varepsilon_t - \theta\varepsilon_{t-1},\ \varepsilon_t \overset{iid}{\sim} N(0,0{,}1)$; $v_t = -\widehat{\theta}\widehat{\varepsilon}_{t-1}$; $\alpha = 5\,\%$, 1000 repeticiones

θ	ρ_{yv}	T	B	H-M	TC 2×2	TC $m\times m$	P-T	C1-v1	C1-v2	C2	C3
0,204	0,2	10	10,2	12,8	11,2	1,4	7,8	2,3	1,9	0,6	0,2
		25	15,3	17,5	12,2	0,6	2,3	1,8	2,1	1,1	1,1
		50	22,8	26,4	20,2	1,4	3,2	2,3	5,5	1,7	2,1
		100	32,9	38,2	28,2	3,5	6,7	2,9	11,8	2,0	4,5
0,314	0,3	10	13,2	16,3	13,3	2,2	9,4	10,2	10,1	5,9	4,7
		25	23,6	26,9	21,1	4,4	10,6	14,8	16,7	10,8	10,1
		50	38,0	40,8	32,1	12,7	14,4	24,4	29,9	19,5	15,8
		100	57,8	62,0	52,0	29,7	24,1	37,0	44,5	33,3	27,7
0,436	0,4	10	20,9	20,2	19,1	3,3	15,9	22,7	21,8	15,3	11,3
		25	38,2	40,1	32,6	13,0	18,2	36,8	36,7	30,7	21,8
		50	58,9	61,8	54,2	35,8	31,5	57,1	57,6	52,5	41,3
		100	85,0	87,0	80,6	70,8	53,9	81,5	82,6	78,1	69,6
0,577	0,5	10	27,6	27,0	24,4	4,9	21,9	36,1	34,2	25,5	21,2
		25	51,0	56,1	42,8	26,1	29,6	58,7	56,7	51,2	39,9
		50	77,7	78,0	72,7	62,3	51,2	82,5	81,5	78,8	68,1
		100	96,4	97,2	95,6	94,5	82,4	97,2	97,2	96,3	93,8
0,750	0,6	10	33,7	31,9	33,3	10,0	28,1	48,3	47,7	35,8	30,5
		25	67,2	68,0	59,8	49,1	49,0	76,7	76,9	72,3	61,3
		50	91,9	93,5	90,6	87,8	77,7	95,0	95,3	94,4	90,2
		100	99,6	99,8	99,5	99,8	97,5	99,8	99,8	99,8	99,3
0,927	0,68	10	41,9	36,7	38,0	13,3	35,0	55,4	53,7	43,1	37,2
		25	79,0	77,5	68,9	64,5	63,1	88,0	87,6	84,8	75,0
		50	97,6	97,7	95,8	96,5	90,9	99,4	99,2	99,0	96,5
		100	100,0	100,0	100,0	100,0	99,9	100,0	100,0	100,0	99,9

Tabla 2. Probabilidad Rechazo Experimento 2 (PGD: MA(1); Modelo Estimado: AR(1))

$y_t = \varepsilon_t - \theta\varepsilon_{t-1},\ \varepsilon_t \overset{iid}{\sim} N(0,0{,}1)$; $v_t = -\widehat{\phi} y_{t-1}$; $\alpha = 5\,\%$, 1000 repeticiones

θ	ρ_{yv}	T	B	H-M	TC 2×2	TC $m\times m$	P-T	C1-v1	C1-v2	C2	C3
$0,209$	$0,2$	10	9,4	13,3	11,5	1,6	8,1	3,3	2,5	1,1	0,6
		25	16,8	20,0	13,5	0,6	3,6	2,3	2,5	1,6	1,9
		50	22,8	26,0	18,5	1,5	3,6	3,0	6,5	1,9	3,0
		100	34,2	39,4	28,9	4,0	6,2	2,9	11,1	2,1	4,5
$0,333$	$0,3$	10	12,6	16,1	14,5	2,6	9,7	13,8	11,4	7,8	6,5
		25	23,7	27,0	19,9	5,5	9,6	16,4	16,8	12,7	11,9
		50	39,3	41,8	33,9	12,0	13,8	24,8	29,7	20,7	16,1
		100	61,3	65,6	54,4	31,0	25,9	41,7	50,4	36,3	30,4
$0,500$	$0,4$	10	20,3	21,5	22,5	4,1	12,5	22,8	18,6	14,4	12,5
		25	38,7	39,8	31,2	15,3	20,6	36,1	33,8	29,7	23,3
		50	58,1	62,7	52,6	34,5	30,1	54,9	54,8	51,4	40,2
		100	83,7	86,2	79,0	69,9	51,5	79,7	79,7	77,3	68,1
$0,905$	$0,49$	10	22,1	22,2	24,9	8,4	18,2	30,9	27,5	22,0	18,2
		25	51,3	52,6	42,9	25,6	28,0	52,0	49,4	47,2	37,0
		50	78,9	79,7	72,5	63,4	49,6	77,5	76,7	75,4	66,1
		100	96,6	97,3	95,8	95,0	80,6	97,0	97,0	96,3	93,0

Tabla 3. Probabilidad Rechazo Experimento 3 (PGD: Lineal x_{1t}, x_{2t}; Mod. Estimado: Lineal x_{1t})

$$y_t = \beta_1 x_{1t} + \beta_2 x_{2t} + \varepsilon_t,\ (x_{1t}, x_{2t}, \varepsilon_t)' \overset{iid}{\sim} N(0_{3\times 1}, \Sigma),\ \Sigma = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \sigma_2^2 & 0 \\ 0 & 0 & 0,1 \end{pmatrix};\ v_t = \widehat{\beta}_1 x_{1t},$$

$\alpha = 5\,\%$, 1000 repeticiones

σ_2^2	ρ_{yv}	T	B	H-M	TC 2×2	TC $m\times m$	P-T	C1-v1	C1-v2	C2	C3
23,9	0,2	10	9,5	8,9	8,5	1,4	8,2	3,7	3,0	0,6	0,2
		25	14,2	14,9	9,8	0,5	2,5	1,7	2,4	1,1	0,9
		50	22,5	21,8	15,9	1,5	3,7	2,3	6,4	1,7	2,8
		100	34,0	35,1	23,9	4,1	7,3	3,0	11,9	2,5	4,8
10,01	0,3	10	14,7	12,1	12,2	3,1	7,7	10,9	11,4	6,6	3,7
		25	24,4	23,9	16,1	4,7	10,7	15,3	18,3	10,3	10,5
		50	39,5	38,1	30,7	11,7	14,8	23,0	29,3	18,6	16,0
		100	61,2	62,4	52,6	27,2	23,8	37,7	47,5	33,6	26,9
2,9	0,5	10	26,7	21,7	22,5	8,4	19,5	32,8	32,3	22,6	19,9
		25	50,8	48,3	40,9	25,9	32,0	55,9	54,9	50,6	39,2
		50	77,8	75,5	69,1	61,0	53,3	81,7	80,3	78,3	67,6
		100	96,3	96,1	93,1	94,0	81,9	96,7	96,9	96,1	93,3
0,94	0,7	10	49,7	40,3	42,2	21,3	43,8	60,0	59,7	50,5	43,2
		25	85,4	81,6	75,2	68,9	69,1	88,3	87,6	85,3	77,8
		50	98,4	98,0	96,8	95,8	92,3	98,9	98,9	98,9	96,8
		100	100,0	100,0	99,9	100,0	99,9	100,0	100,0	100,0	100,0
0,134	0,9	10	81,3	70,4	74,7	57,6	76,7	89,3	89,0	84,2	76,3
		25	99,3	98,4	97,9	98,9	98,9	99,8	99,6	99,7	99,4
		50	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
		100	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
0,008	0,95	10	88,5	77,0	80,4	75,8	89,2	95,6	95,5	92,4	89,4
		25	100,0	99,9	99,6	100,0	100,0	100,0	100,0	100,0	100,0
		50	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
		100	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0

Tabla 4. Probabilidad Rechazo Experimento 4 (Datos y Prevs distribuidos por Normal Biv.)
$(y_t, v_t) \sim N(0_{2\times 1}, \Sigma)$, siendo $\Sigma = \begin{pmatrix} \sigma_y^2 & \rho\sigma_y\sigma_v \\ \rho\sigma_y\sigma_v & \sigma_v^2 \end{pmatrix}$, $\sigma_y^2 = \sigma_v^2 = 1$,
$\alpha = 5\,\%$, 1000 repeticiones

$\rho_{yv} = \rho$	T	B	H-M	TC 2×2	TC $m \times m$	P-T	C1-v1	C1-v2	C2	C3
0,0	10	5,4	4,5	7,8	2,9	12,3	5,8	8,0	4,3	3,4
	25	5,0	5,4	5,8	4,7	6,9	4,7	5,2	4,3	2,9
	50	5,4	5,5	5,3	5,5	6,6	4,8	5,9	5,0	2,8
	100	4,5	5,9	5,4	5,6	5,5	5,0	5,5	4,8	3,3
0,2	10	10,5	10,3	9,9	4,4	13,7	11,2	13,6	9,1	6,8
	25	17,5	17,0	12,6	8,4	11,0	15,7	17,7	14,2	8,9
	50	23,1	23,3	17,5	11,2	11,6	22,2	24,3	20,6	13,8
	100	34,7	37,2	26,8	20,2	21,8	38,1	40,1	37,1	26,7
0,3	10	14,5	14,3	13,1	5,5	15,5	17,9	22,8	13,2	9,1
	25	25,6	24,8	18,3	11,9	16,0	27,0	27,6	24,2	17,8
	50	39,3	38,3	28,3	21,8	24,2	42,3	44,9	41,0	28,6
	100	59,3	60,3	49,7	42,6	43,2	66,9	68,1	65,9	49,6
0,4	10	18,8	17,0	14,5	6,7	18,9	24,3	28,6	17,7	13,7
	25	37,3	33,8	25,8	18,8	25,2	42,0	44,0	39,4	27,8
	50	58,1	56,1	48,9	42,0	42,6	65,1	67,3	63,9	51,5
	100	84,5	85,2	77,6	74,5	70,4	90,0	90,9	89,5	80,8
0,5	10	27,9	22,3	22,2	10,9	21,3	29,2	33,3	23,0	19,5
	25	53,8	48,7	41,3	29,5	36,3	58,6	60,3	54,9	40,9
	50	77,7	75,5	69,2	62,5	62,6	82,9	84,5	81,7	71,1
	100	96,3	96,0	93,7	93,6	89,5	97,5	97,8	97,4	94,7
0,7	10	47,8	37,9	38,5	21,8	39,3	54,6	58,1	46,0	37,9
	25	79,3	77,1	71,0	66,4	67,9	87,8	88,9	85,5	74,5
	50	97,5	97,0	95,7	96,6	93,6	99,4	99,3	99,0	97,0
	100	99,9	99,9	99,8	100,0	100,0	100,0	100,0	100,0	100,0
0,9	10	80,1	66,4	72,9	54,9	75,8	88,2	88,5	82,2	74,2
	25	99,5	98,5	98,0	99,3	98,5	99,9	100,0	99,9	99,2
	50	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	100	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
0,95	10	90,1	79,8	82,9	75,7	89,8	95,8	95,8	93,9	89,8
	25	99,8	99,8	99,7	99,9	99,9	99,9	100,0	100,0	99,9
	50	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	100	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
−0,25	10	1,6	1,8	11,5	3,5	18,9	1,7	2,2	1,0	1,4
	25	0,5	0,8	13,2	9,5	17,1	0,7	1,0	0,7	0,8
	50	0,2	0,2	22,6	14,9	22,1	0,2	0,3	0,2	0,1
	100	0,1	0,1	37,2	31,5	31,0	0,0	0,1	0,0	0,1
−0,8	10	0,0	0,2	51,6	32,1	30,4	0,0	0,2	0,0	0,7
	25	0,0	0,0	87,7	88,5	49,7	0,0	0,0	0,0	0,3
	50	0,0	0,0	99,3	99,9	78,4	0,0	0,0	0,0	0,2
	100	0,0	0,0	100,0	100,0	96,9	0,0	0,0	0,0	0,1

Tabla 5. Probabilidad Rechazo Experimento 5 (Datos y Prevs Normal Biv, σ_v^2 varía)
$(y_t, v_t) \sim N(0_{2\times 1}, \Sigma)$, siendo $\Sigma = \begin{pmatrix} \sigma_y^2 & \rho\sigma_y\sigma_v \\ \rho\sigma_y\sigma_v & \sigma_v^2 \end{pmatrix}$, $\sigma_y^2 = 1$, $\rho = 0,7$,
$\alpha = 5\,\%$, 1000 repeticiones

σ_v^2	T	B	H-M	TC 2×2	TC $m\times m$	P-T	C1-v1	C1-v2	C2	C3
1,0	10	49,1	39,9	39,5	21,2	41,6	56,3	58,0	47,9	38,3
	25	83,3	79,0	73,0	70,1	72,5	90,2	90,2	88,5	80,9
	50	98,0	97,4	96,4	97,7	94,9	99,3	99,3	99,3	98,0
	100	99,9	99,9	99,8	100,0	100,0	100,0	100,0	100,0	100,0
0,1	10	46,5	37,8	40,5	6,6	22,7	38,8	36,6	24,3	18,6
	25	82,4	78,0	72,5	33,0	52,2	68,9	70,5	60,9	58,0
	50	97,2	97,2	95,1	81,0	78,9	91,4	93,0	88,5	85,7
	100	100,0	100,0	100,0	99,1	96,8	99,0	99,4	98,8	99,0

5.2.2. Decisiones opuestas de los tests

Como ya se ha venido comentando en el apartado anterior, las diferencias entre el comportamiento de unos tests y otros son a veces difíciles de valorar si nos basamos simplemente en comparar sus probabilidades de rechazo. Esto ocurre especialmente cuando ρ_{yv} toma un valor positivo pero no muy elevado (si es muy elevado, es obvio que los tests deberían rechazar la hipótesis nula y si es negativo, no rechazarla). Los tests C1-v1, C1-v2 y C2 tienden a presentar una probabilidad de rechazo más alta que los de la literatura en dichos casos, pero, a priori, no sabemos si esto es o no deseable. Por ello, a lo largo de la ejecución de los experimentos de simulación anteriores, hemos ido recopilando todos los casos en los que los contrastes de la literatura y los nuestros deciden de forma opuesta (unos no rechazan y otros rechazan, y viceversa), para exponer un resumen de ellos ahora, y que el propio lector pueda valorar cuál de las dos decisiones le parece más correcta o razonable en cada caso, y, de este modo, comprobar si nuestros tests parecen ser ventajosos en la práctica.

Las variables que interesan en este contexto para caracterizar una muestra de datos y previsiones son las frecuencias relativas asociadas a cada posible valor de la función de pérdidas, es decir, la estimación del vector p definido en el apartado 3.2. Dicho de otro modo, nos interesa saber qué porcentaje de las previsiones generaron pérdida 0, 1, 2, etc, y luego juzgar si la decisión tomada por los contrastes de la literatura es razonable, dados dichos valores, o lo es la tomada por los nuestros. Para complementar dicha información, también interesa la estimación del vector q definido en (4), es decir, interesa conocer las frecuencias relativas que se hubieran generado por la muestra si la hipótesis de independencia estocástica entre datos y previsiones fuera cierta. También conviene resumir toda la información de los vectores $\widehat{p}$ y $\widehat{q}$ a través de la media muestral de pérdidas observada y la obtenida bajo independencia entre datos y previsiones, $a\widehat{p}$ y $a\widehat{q}$ (recuérdese que a es el vector fila con los valores de la función de pérdida y que $a\widehat{p}$ y $a\widehat{q}$ son los estadísticos fundamentales para la definición de los contrastes C1-v1, C1-v2 y C2).

Hemos elegido mostrar los resultados generados en los experimentos cuando se utilizó la partición de cuatro regiones y función de pérdida (16), porque la interpretación de las pérdidas bajo esa partición es más intuitiva que con partición de tres regiones y función (15), ya que, con la primera, se evalúa el acierto de las previsiones respecto a signo y magnitud. De nuevo, hemos decidido centrarnos en los resultados obtenidos en el Experimento 4, para evitar redundancias.[30]

Como ya se mencionó más arriba, el criterio para seleccionar una simulación para su análisis ha sido que los tests C1-v1, C1-v2 y C2 tomaran una decisión para ella, mientras los tests TC $m\times m$ y P-T toman la decisión contraria. Solo nos interesan los contrastes que particionan el dominio de datos y previsiones en m regiones, que son los verdaderos competidores de los tests que presentamos en este documento, eliminando los que usan solo dos regiones, que solo evaluarían el signo de las previsiones. Respecto a los nuestros, hemos excluido a C3, porque actúa de modo bastante similar a P-T. Hemos separado los casos en que los

[30] Por supuesto, se han generado para el resto de experimentos las mismas tablas que aquí se van a mostrar para el Experimento 4, y las conclusiones son las mismas.

tres tests de la literatura no rechazaban la hipótesis nula (previsiones no útiles) mientras los propuestos en el documento sí la rechazaban (previsiones útiles), y aquellos en los que ocurría lo contrario. Denotaremos a los primeros por simulaciones tipo R y por simulaciones tipo NR a los segundos.[31] Éstas últimas son muy poco frecuentes, salvo en ejercicios donde $\rho_{yv} < 0$, pero este caso ya había sido identificado, y fue ampliamente comentado en el apartado anterior y en 3.1, motivando originalmente la creación de nuestros tests. Las simulaciones tipo R sí se producen con gran frecuencia, aunque menos cuanto mayores son ρ_{yv} y el tamaño muestral T.

Pues bien, adjuntamos en el Apéndice F los datos de las simulaciones de interés para cada uno de los diez diseños del Experimento 4, y, a su vez, para cada uno de los cuatro tamaños muestrales. De las 1000 ejecuciones de cada uno de esos 40 ejercicios, existen una gran cantidad de ellas en las que los tests citados deciden de forma opuesta (denotemos por *num* dicha cantidad). De todas ellas, se han elegido para mostrar en las tablas solo tres simulaciones representativas: aquellas en las que la diferencia absoluta $|a\widehat{p} - a\widehat{q}|$ fue la máxima de entre las *num* simulaciones de interés, aquellas en las que dicha diferencia absoluta fue mínima y, además, aquella que se corresponde con la mediana de la serie de los *num* valores absolutos de las diferencias. Para cada uno de los 40 ejercicios, la información que se presenta en las tablas del Apéndice F es la siguiente:

a) *num*: número de simulaciones tipo R (NR), de las 1000 ejecutadas para el ejercicio.

Luego, para cada una de las tres simulaciones elegidas de las *num* anteriores (máximo, mínimo y mediana de las diferencias $|a\widehat{p} - a\widehat{q}|$):

b) Valor numérico de la media muestral de las pérdidas ($a\widehat{p}$) y de la media que se obtendría bajo independencia ($a\widehat{q}$).

c) Valor numérico de las estimaciones $\widehat{p}$ y $\widehat{q}$. En la notación utilizada a lo largo del documento, p_i representa la probabilidad de que la pérdida z_t asociada a un par (y_t, v_t) tome valor a_i. Para facilitar la interpretación de resultados, en los cuadros a continuación se utilizará, para referirnos a dicho concepto, la notación $p(a_i)$. Como la función de pérdidas usada es (16), siendo las regiones r_1, r_2, r_3, r_4 de la partición las descritas en el apartado 5.1.3, las pérdidas posibles son 0, 1, 2 y 3, y su significado es el siguiente: 0, acierto del signo y magnitud del dato; 1, acierto del signo pero imprecisión en magnitud; 2, error en la previsión del signo, no siendo la magnitud del error la máxima posible (es decir, no se asignó la previsión en el cuadrante más lejano posible del correcto, sino en el contiguo); 3, error en la previsión del signo, y máxima magnitud de error de previsión posible. Así que $\widehat{p}(k)$ es la estimación de la probabilidad correspondiente a la pérdida de valor k, y, análogamente, $\widehat{q}(k)$ designará la estimación de dicha probabilidad pero bajo el supuesto de independencia entre datos y previsiones.

Es inviable presentar en este apartado todos los resultados anteriores, además de redundante. Por ello, expondremos aquí la información de las simulaciones tipo R y NR correspondiente a uno solo de los diez diseños del Experimento 4 asociados a cada tamaño muestral ($T = 10, 25, 50, 100$). Para construir los Cuadros 8 y 9 hemos elegido, para cada T, el diseño con el que se produjeron las simulaciones tipo R (NR) con mayor $|a\widehat{p} - a\widehat{q}|$ de los diez considerados. Por tanto, de los 40 ejercicios posibles, solo seleccionamos en los cuadros a continuación los resultados de cuatro (uno para cada valor de T), y, de éstos, solo mostramos las tres simulaciones tipo R (NR) representativas. No obstante, como puede observarse en las tablas del Apéndice F, los resultados de cualquier diseño conducirían a las mismas conclusiones cualitativas.

[31] La letra pretende denotar la decisión de los tests C1-v1, C1-v2 y C2: R, si rechazaban; NR, si no rechazaban.

Cuadro 8. Caracterización Simulaciones tipo R representativas

	$a\widehat{p}$	$a\widehat{q}$	$\widehat{p}(0)$	$\widehat{p}(1)$	$\widehat{p}(2)$	$\widehat{p}(3)$	$\widehat{q}(0)$	$\widehat{q}(1)$	$\widehat{q}(2)$	$\widehat{q}(3)$
	$T = 10$		Diseño: 6 ($\rho = 0,7$)				$num = 149$ (sobre 1000)			
Máxima	0,50	1,62	0,50	0,50	0	0	0,26	0,26	0,08	0,40
Mediana	0,80	1,54	0,50	0,20	0,30	0	0,26	0,24	0,20	0,30
Mínima	0,90	1,40	0,20	0,70	0,10	0	0,19	0,35	0,33	0,13
	$T = 25$		Diseño: 4 ($\rho = 0,4$)				$num = 173$ (sobre 1000)			
Máxima	0,84	1,56	0,44	0,36	0,12	0,08	0,29	0,20	0,15	0,36
Mediana	1,04	1,51	0,48	0,24	0,04	0,24	0,32	0,19	0,14	0,35
Mínima	1,16	1,51	0,28	0,40	0,20	0,12	0,23	0,26	0,29	0,22
	$T = 50$		Diseño: 4 ($\rho = 0,4$)				$num = 185$ (sobre 1000)			
Máxima	0,98	1,55	0,42	0,34	0,08	0,16	0,30	0,20	0,15	0,35
Mediana	1,18	1,52	0,36	0,28	0,18	0,18	0,29	0,22	0,18	0,31
Mínima	1,26	1,52	0,32	0,28	0,22	0,18	0,26	0,24	0,22	0,28
	$T = 100$		Diseño: 3 ($\rho = 0,3$)				$num = 178$ (sobre 1000)			
Máxima	1,14	1,51	0,37	0,30	0,15	0,18	0,29	0,22	0,18	0,31
Mediana	1,33	1,58	0,36	0,23	0,13	0,28	0,29	0,21	0,14	0,36
Mínima	1,36	1,56	0,33	0,24	0,17	0,26	0,27	0,23	0,16	0,33

Las simulaciones se toman del Expmto 4, con partición de 4 regiones/función de 4 pérdidas

Cuadro 9. Caracterización Simulaciones tipo NR representativas

	$a\widehat{p}$	$a\widehat{q}$	$\widehat{p}(0)$	$\widehat{p}(1)$	$\widehat{p}(2)$	$\widehat{p}(3)$	$\widehat{q}(0)$	$\widehat{q}(1)$	$\widehat{q}(2)$	$\widehat{q}(3)$
	$T = 10$		Diseño: 4 ($\rho = 0,4$)				$num = 3$ (sobre 1000)			
Máxima	2,10	1,48	0	0,40	0,10	0,50	0,26	0,26	0,22	026
Mediana	1,10	1,53	0,50	0,10	0,20	0,20	0,30	0,20	0,17	0,33
Mínima	1,10	1,53	0,50	0,10	0,20	0,20	0,30	0,20	0,17	0,33
	$T = 25$		Diseño: 2 ($\rho = 0,2$)				$num = 2$ (sobre 1000)			
Máxima	2,12	1,52	0,12	0,20	0,12	0,56	0,32	0,18	0,16	0,34
Mediana	1,72	1,54	0,08	0,36	0,32	0,24	0,25	0,25	0,20	0,30
Mínima	1,72	1,54	0,08	0,36	0,32	0,24	0,25	0,25	0,20	0,30
	$T = 50$		Diseño: 1 ($\rho = 0,0$)				$num = 9$ (sobre 1000)			
Máxima	2,22	1,54	0,08	0,16	0,22	0,54	0,28	0,22	0,18	0,32
Mediana	1,74	1,47	0,20	0,26	0,14	0,40	0,32	0,21	0,15	0,32
Mínima	1,64	1,52	0,16	0,32	0,24	0,28	0,27	0,24	0,18	,31
	$T = 100$		Diseño: 1 ($\rho = 0,0$)				$num = 6$ (sobre 1000)			
Máxima	2,04	1,55	0,18	0,12	0,18	0,52	0,28	0,23	0,17	0,33
Mediana	1,73	1,54	0,20	0,27	0,13	0,40	0,30	0,20	0,15	0,34
Mínima	1,48	1,57	0,39	0,10	0,15	0,36	0,29	0,21	0,14	0,36

Las simulaciones se toman del Expmto 4, con partición de 4 regiones/función de 4 pérdidas

Análisis de Simulaciones tipo R Los resultados del Cuadro 8 sugieren que, en prácticamente todos aquellos casos en los que los contrastes C1-v1, C1-v2 y C2 están rechazando la hipótesis nula (en favor de la utilidad de las previsiones), mientras TC $m \times m$ y P-T resolvían lo contrario, son los primeros los que parecen estar actuando correctamente. Como ya habíamos apuntado arriba, este tipo de situaciones se presentan con mucha mayor frecuencia cuando ni ρ_{yv} ni T son muy elevados.

Los resultados correspondientes a longitud muestral $T = 10$ son indiscutibles. Véase la primera de las tres simulaciones presentadas en el Cuadro 8 para dicha longitud. Las previsiones son correctas en signo en el 100 % de los casos ($\widehat{p}(0) + \widehat{p}(1) = 1$), acertando la magnitud del dato además el 50 % de las ocasiones ($\widehat{p}(0) = 0,5$), y, sin embargo, los tests TC $m \times m$ y P-T mantienen la hipótesis de que las previsiones no son útiles. Los nuestros, la rechazan, tal y como parece obvio que debería ser. De entre las 149 simulaciones en las que nuestros tests rechazan y los dos elegidos de la literatura no rechazan, ésta es en la que se produce un error de éstos últimos más claro. Pero si tomamos la simulación correspondiente a

la mediana respecto a $|a\widehat{p} - a\widehat{q}|$ de esas 149 simulaciones tipo R (segunda fila del caso $T = 10$ del Cuadro 8), también la decisión de nuestros contrastes es mucho más razonable. Las previsiones habían acertado el signo del dato en un 70 % de los casos ($\widehat{p}(0) + \widehat{p}(1) = 0,7$), y, de ellos, el 50 % de las veces habían sido, además, exactas en magnitud ($\widehat{p}(0) = 0,5$). Incluso en el caso que podría ser más desfavorable para juzgar la actuación de nuestros tests (es decir, aquel en el que la decisión de nuestros tests fue menos rotunda) de las 149 simulaciones tipo R correspondientes a esta longitud muestral (tercera fila del caso $T = 10$ del Cuadro 8), a nuestro juicio, también parece más adecuado decidir como lo hacen C1-v1, C1-v2 y C2 que como los otros: las previsiones han sido certeras en signo el 90 % de las veces ($\widehat{p}(0) + \widehat{p}(1) = 0,9$), aunque solo el 20 % de ellas hayan sido correctas en magnitud ($\widehat{p}(0) = 0,2$). Parece razonable rechazar la conjetura de que las previsiones no tienen utilidad.

El mismo análisis podría efectuarse para el resto de longitudes muestrales. Como ya hemos dicho, la superioridad de nuestros contrastes es menos clara al aumentar T, pero creemos que sigue existiendo. Para verlo, de las tres simulaciones presentadas para cada T, concentrémonos siempre en la simulación tipo R correspondiente a la mediana de la serie $|a\widehat{p} - a\widehat{q}|$ (Cuadro 8). El porcentaje de veces que las previsiones fueron correctas en signo fue el 72 %, 64 % y 59 % (valores obtenidos de las sumas $\widehat{p}(0)+\widehat{p}(1)$), para muestras de tamaño 25, 50 y 100, respectivamente. Además, el 48 %, 36 % y 36 % de las veces, para $T = 25$, 50 y 100, respectivamente, la previsión fue completamente certera, es decir, anticipó correctamente la magnitud del dato, además del signo (valores de $\widehat{p}(0)$). Parecen resultados suficientemente buenos como para rechazar que las previsiones carezcan de valor. El 17,3 %, 18,5 % y 17,8 % de las simulaciones del Experimento 4 (con partición de 4 regiones y función de pérdida (16)) para $T = 25$, 50 y 100, respectivamente, son similares a las que se acaban de analizar (los valores resultan del ratio entre la variable *num* del Cuadro 8 y el número de repeticiones del ejercicio, 1000) y, sin embargo, TC $m \times m$ y P-T mantienen ambos la ausencia de valía de las previsiones. En cambio, C1-v1, C1-v2 y C2 rechazan dicha hipótesis, decisión que, a nuestro juicio, es más adecuada.

El lector podría pensar que quizá los tests C1-v1, C1-v2 y C2 también están rechazando la hipótesis nula en situaciones donde es claro que no deberían. Para chequear si esto es así, se han presentado en el Cuadro las simulaciones tipo R correspondientes al mínimo de la serie $|a\widehat{p} - a\widehat{q}|$. Según éstas, en ninguno de los cuatro tamaños muestrales considerados los tests citados cometen un error claro. Solo en el caso $T = 100$, la información de la simulación tipo R con menor $|a\widehat{p} - a\widehat{q}|$ indica que podrían existir simulaciones donde la decisión de nuestros tests sería discutible.

Análisis de Simulaciones tipo NR Los valores de la variable *num* presentandos en el Cuadro 9 sugieren que, cuando nuestros tests no rechazan la hipótesis nula de no utilidad de las previsiones, tampoco lo hacen los contrastes ya existentes en la literatura. Por supuesto, la excepción a esto son los escenarios donde la correlación entre datos y previsiones era negativa, tal y como se documentó en el apartado 5.2.1 (estos casos no se han incluido en el Cuadro 9 puesto que el erróneo comportamiento de los tests TC $m \times m$ y P-T en estas situaciones ya era conocido). Fuera de estas circunstancias, es muy improbable encontrar simulaciones tipo NR. Se producen con una frecuencia menor al 1 % en diseños de ρ_{yv} no negativa, como puede comprobarse a través de la variable *num* en las tablas del Apéndice F. En el Cuadro 9 hemos adjuntado algunos de los pocos casos en que esto ocurrió en nuestras ejecuciones del Experimento 4, y sugieren que, desde luego, cuando nuestros contrastes no rechazan la hipótesis nula a la vez que los otros tests sí lo hacen, son los primeros los que están actuando correctamente. Solo en alguna de las simulaciones tipo NR correspondientes a la longitud $T = 10$ la decisión correcta sobre la hipótesis nula es discutible. En el resto, el porcentaje de aciertos en signo de las previsiones es inferior al 50 % ($\widehat{p}(0) + \widehat{p}(1) < 0,5$) y, además, de las cuatro pérdidas posibles, la que ocurre con mayor frecuencia es la correspondiente a error en signo y máximo error en magnitud, pérdida de valor numérico 3 (es decir, $\widehat{p}(3) > \widehat{p}(k)$, para $k = 0, 1, 2$). Sin embargo, los tests TC $m \times m$ y P-T estaban rechazando la hipótesis nula en favor de la utilidad de las previsiones; los nuestros mantienen la hipótesis nula, tal y como parece razonable.

Conclusiones En el apartado 5.2.1 quedó claro que nuestros tests se comportaban mejor que los contrastes estándar cuando la correlación entre datos y previsiones ρ_{yv} es negativa (caso en el que debe no rechazarse la hipótesis nula) y cuando es positiva y muy elevada (caso en el que sí debe rechazarse). El escenario en el que no está claro a priori cuál es el comportamiento deseable de los tests que contrastan sobre la utilidad de previsiones era aquel en el que ρ_{yv} es positivo pero no muy elevado, digamos cuando

$0 < \rho_{yv} < 0,7$. Pues bien, los resultados presentados en este apartado aclaran aquella duda: también en dichas situaciones, los contrastes que se han presentado en este documento deciden de forma más *razonable* que aquellos tests estándar de la literatura que permiten particionar el dominio de datos y previsiones en un número de regiones $m \geq 2$.

6. Conclusiones

Los tests estadísticos para valorar si un conjunto de previsiones es útil presentan el problema general de la subjetividad sobre la definición de utilidad de la predicción y la dificultad de formalizar dicho concepto. Los contrastes utilizados habitualmente en la literatura particionan el dominio de los datos en m regiones, y valoran las previsiones según su capacidad de predecir la región en que se situarán los datos. Estos tests se clasifican en dos grupos, según restrinjan la partición a solo dos regiones o no. Los que así lo hacen (H-M, B y TC 2×2), reducen el concepto de utilidad de las previsiones esencialmente a su capacidad de predecir correctamente el signo del dato,[32] enfoque claramente insuficiente en la mayoría de aplicaciones. Los que admiten particiones más finas (TC $m \times m$ y P-T) identifican previsiones útiles con dependencia estocástica entre éstas y los datos, y, sorprendentemente, no introducen una función de pérdidas en su especificación. Dichas cuestiones provocan que existan muchas situaciones en las que estos tests deciden incorrectamente sobre la utilidad de un conjunto de previsiones. Nuestra propuesta es esencialmente introducir una función de pérdidas de naturaleza "discreta", que asigna penalizaciones a los m^2 cuadrantes en que quedará particionado el dominio bidimensional de los datos y las previsiones. Esto permitirá a) que en la valoración de las previsiones intervengan, en general, tanto el acierto en signo como en magnitud y b) que el concepto de utilidad de las previsiones sea definido por el propio usuario, según sus juicios de valor y la aplicación concreta en la que se encuentre, ya que es el usuario del test quien define no solo la partición sino también los valores numéricos de las penalizaciones de los cuadrantes. En este contexto, proponemos dos tipos de contrastes. En primer lugar, un contraste sobre la posición (identificada por la media poblacional) de la distribución de las pérdidas generadas, basado en la media muestral de éstas. Presentamos tres versiones alternativas, las dos primeras (C1-v1 y C1-v2) con validez asintótica, y una tercera (C2), usando una distribución muy próxima a la verdadera en muestras finitas. El carácter discreto de la función de pérdidas que definimos es clave para que la derivación de las distribuciones de los estadísticos de contraste involucrados sea posible. En segundo lugar, un contraste paramétrico especial de Razón de Verosimilitudes (también de validez asintótica), adaptado de una propuesta original de Robertson y Wright (1981), C3. Éste tiene el inconveniente de no tener en cuenta los valores numéricos concretos de la función de pérdidas, por lo que la afirmación b) no aplica con exactitud en su caso.

Hemos realizado pruebas de simulación para chequear las propiedades de los cinco contrastes habituales de la literatura y de los cuatro que aquí se proponen. La ambigüedad de la hipótesis de utilidad de las previsiones no permite un análisis estándar de tamaño y potencia. En cualquier caso, los resultados de los experimentos son muy favorables hacia nuestros tests. Por un lado, no cometen errores inadmisibles en ningún caso (P-T y TC $m \times m$ los cometen cuando la correlación entre datos y previsiones ρ_{yv} es negativa, mientras los tests que definen particiones con $m = 2$ (H-M, B y TC 2×2) los cometen al valorar por igual dos conjuntos de previsiones con tal que su capacidad de predecir el signo sea la misma). Por otro lado, son más potentes en casos en que las previsiones son claramente útiles (lo que sucede cuando ρ_{yv} es positiva y muy elevada) − salvo C3, cuyos resultados son similares a los de P-T −. Finalmente, en situaciones donde la correlación es intermedia, deciden de un modo mucho más "razonable" que P-T y TC $m \times m$, en el sentido de que su decisión es sensible a la definición de utilidad que proporcione el usuario.

Desde nuestro punto de vista, el contraste sobre la esperanza matemática de las pérdidas (en cualquiera de sus tres versiones, C1-v1, C1-v2 y C2) es preferible al test de Razón de Verosimilitudes, C3, por razones teóricas y empíricas. Sobre la versión del primer tipo de contraste que aconsejamos, podemos decir que la derivación teórica de C1-v2 es más correcta, pero sus propiedades en muestras finitas no difieren mucho de las de C1-v1 y, por el contrario, su cálculo es algo más complejo. Por su parte, C2 presenta un coste computacional muy elevado, y, en la práctica, no se han apreciado mejores resultados para este contraste en muestras muy cortas.[33] Finalmente, aunque la derivación y análisis de los tests se ha realizado bajo el supuesto de no existencia de incertidumbre paramétrica asociada a las previsiones, los resultados del Capítulo 3 de esta Tesis sugieren que los contrastes C1-v1 y C1-v2 son robustos a la presencia de ésta en la mayor parte de contextos de previsión habituales en la práctica, sin necesitar corrección alguna respecto a la especificación con la que aparecen en este capítulo, aunque las predicciones hubieran sido generadas a partir de modelos estimados.

[32] O el signo de la diferencia entre el dato y el valor usado para generar la partición.

[33] Debido a ser un contraste que aplica una distribución que, en teoría, debe ser más próxima a la verdadera que la asintótica, se podrían haber esperado ventajas respecto a los otros tests cuando las muestras fueran cortas.

A. Apéndice: Cálculos necesarios para la aplicación de los contrastes propuestos

Se presentan en este apéndice algunos cálculos o procedimientos necesarios para la implementación de los contrastes, que quedaron pendientes cuando se realizó la exposición teórica de éstos.

A.1. Cálculos requeridos para la aplicación del contraste C1-v2. Cálculo de $\nabla f(P)$

Recuérdese que la función f es

$f(P) = p - q = \left(\sum_{(r_i,r_j)\in R_1} p_{ij} - (p_i^y p_j^v), \sum_{(r_i,r_j)\in R_2} p_{ij} - (p_i^y p_j^v), ..., \sum_{(r_i,r_j)\in R_J} p_{ij} - (p_i^y p_j^v) \right)'$ y que R_s es el conjunto de cuadrantes (r_k, r_q) cuya pérdida asignada es a_s, es decir, $R_s = \{(r_k, r_q) | g(r_k, r_q) = a_s\}$.

Pues bien, definamos ahora los siguientes conjuntos:

$R_s^{y,i} = \{(r_i, r_q) | (r_i, r_q) \in R_s\}$. Es decir, es el conjunto de cuadrantes de R_s correspondientes a la fila i (por tanto, cuadrantes de la fila i con pérdida asignada a_s).

$R_s^{v,j} = \{(r_k, r_j) | (r_k, r_j) \in R_s\}$. Es decir, es el conjunto de cuadrantes de R_s correspondientes a la columna j (por tanto, cuadrantes de la fila j con pérdida asignada a_s).

La matriz que se requiere calcular es la matriz $J \times m^2$ $\nabla f(P) = \begin{pmatrix} \frac{\partial f_1}{\partial p_{11}} & \frac{\partial f_1}{\partial p_{12}} & \cdots & \frac{\partial f_1}{\partial p_{mm}} \\ \frac{\partial f_2}{\partial p_{11}} & \frac{\partial f_2}{\partial p_{12}} & \cdots & \frac{\partial f_2}{\partial p_{mm}} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial f_J}{\partial p_{11}} & \frac{\partial f_J}{\partial p_{12}} & \cdots & \frac{\partial f_J}{\partial p_{mm}} \end{pmatrix}$, siendo f_s la componente s-ésima de $f(P)$, es decir, $f_s = \sum_{(r_k,r_q)\in R_s} p_{kq} - (p_k^y p_q^v)$.

Antes de dar la expresión general del elemento $\frac{\partial f_s}{\partial p_{ij}}$, vamos a resolver un caso particular, que nos proporcionará la intuición adecuada para entender luego dicha expresión:

Supongamos la siguiente función de pérdida 4×4 (ya utilizada anteriormente a lo largo del documento):

		v_t			
		r_1	r_2	r_3	r_4
y_t	r_1	0	1	2	3
	r_2	1	0	2	3
	r_3	3	2	0	1
	r_4	3	2	1	0

Por ejemplo, razonemos la obtención de la derivada $\frac{\partial f_2}{\partial p_{31}}$:

La función es $f_2 = \sum_{(r_k,r_q)\in R_2} p_{kq} - (p_k^y p_q^v)$. El conjunto R_2 está formado por los cuadrantes con pérdida $a_2 = 1$, es decir, $R_2 = \{(1,2), (2,1), (3,4), (4,3)\}$. Por lo tanto, f_2 toma la expresión $f_2 = [p_{12} - (p_1^y p_2^v)] + [p_{21} - (p_2^y p_1^v)] + [p_{34} - (p_3^y p_4^v)] + [p_{43} - (p_4^y p_3^v)]$.

Debemos encontrar los sumandos donde aparece el parámetro p_{31}. Como las probabilidades marginales p_i^y y p_j^v son la suma de las probabilidades de los cuadrantes de la fila i y de la columna j, respectivamente, el parámetro p_{31} aparece implícitamente en p_3^y y p_1^v. Veáse que los conjuntos $R_2^{y,3}$ y $R_2^{v,1}$ son $R_2^{y,3} = \{(3,4)\}$ y $R_2^{v,1} = \{(2,1)\}$. Como p_{31} aparece en p_3^y y p_1^v, la derivada $\frac{\partial f_2}{\partial P_{31}}$ es $\frac{\partial f_2}{\partial P_{31}} = d_{31}^2 = -p_2^y - p_4^v$. Si el cuadrante $(3,1)$ también hubiera sido de los que se incluyen en R_2, p_{31} sería el primer elemento de uno de los términos entre corchetes, y la derivada sería $1 - d_{31}^2$.

Ahora ya estamos en condición de presentar la expresión de $\frac{\partial f_s}{\partial p_{ij}}$, que será la siguiente:

$\frac{\partial f_s}{\partial p_{ij}} = d_{ij}^s = -\left(\sum_{(r_i,r_q)\in R_s^{y,i}} p_q^v + \sum_{(r_k,r_j)\in R_s^{v,j}} p_k^y \right)$, si $(r_i, r_j) \notin R_s$ y

$\frac{\partial f_s}{\partial p_{ij}} = 1 - d_{ij}^s$, si $(r_i, r_j) \in R_s$.

A.2. Cálculos requeridos para la aplicación del contraste C3

A.2.1. Algoritmo LSA

Sean w vector $r \times 1$ y x vector aleatorio también $r \times 1$ y $C = \{x \in R^r | x_1 \geq x_2 \geq ... \geq x_r\}$, la proyección $E_w(x|C)$ (vector $r \times 1$) puede calcularse a través de un método sencillo (LSA), a saber:

1) Sea el conjunto $Q = \{1, 2, ..., r\}$ y sea $A \subseteq Q$.

2) Se define $M(A) = \frac{\sum_{i \in A} w_i x_i}{\sum_{i \in A} w_i}$.

3) Se obtienen l enteros $(i_0, i_1, ..., i_l)$, tales que $i_0 < i_1 < ... < i_l$, $l \leq r$, del siguiente modo:

3.a) $i_0 = 0$ e $i_l = r$.

3.b) i_1 será el entero $i \leq r$ (si hay varios, el mayor) que maximiza $M(\{i_0 + 1, ..., i\})$, i_2 es el entero i que maximiza $M(\{i_1 + 1, ..., i\})$ y así sucesivamente. El proceso concluye al obtener $i_l = r$.

3.c) Se definen entonces los conjuntos ("level sets") $A_j = \{i_{j-1} + 1, ..., i_j\}$, para $j = 1, 2, ..., l$.

4) La componente iésima de la proyección, es decir, $E_w^{(i)}(x|C)$, es:

$E_w^{(i)}(x|C) = M(A_j)$, si $i \in A_j$, para $j = 1, 2, ..., l$.

Por construcción, $\bigcup_{j=1}^{l} A_j = Q$ y $\bigcap_{j=1}^{l} A_j = \varnothing$. Por lo tanto, la expresión anterior asigna un valor y solo uno a cada una de las r componentes de $E_w(x|C)$.

Tomando $w = \widehat{p}$ y $x = (\widehat{q}_1/\widehat{p}_1, ..., \widehat{q}_J/\widehat{p}_J)$, la proyección $E_w(x|C)$ generada por LSA se aplica, coordenada a coordenada, sobre la estimación $\widehat{p}$ produciendo la estimación $\overline{p}$ necesaria para el estadístico de contraste de C3 (10). Lo primero que debemos tener en cuenta es que la expresión de $M(A)$ siempre va a ser de la forma $\frac{\widehat{q}_1 + ... + \widehat{q}_i}{\widehat{p}_1 + ... + \widehat{p}_i}$, ya que $w_j x_j = \widehat{p}_j \frac{\widehat{q}_j}{\widehat{p}_j} = \widehat{q}_j$. Ilustremos ahora de forma completa cómo funciona el algoritmo LSA y cuál es su papel en el cálculo del vector $\overline{p}$.

Para ello, usaremos el ejemplo expuesto a final del apartado 3.3 que sirvió entonces para aclarar el significado de la hipótesis (5b). Véanse en el cuadro abajo las estimaciones $\widehat{p}$ y $\widehat{q}$ de partida. Los pasos de LSA, tomando $w = \widehat{p}$ y $x = (\widehat{q}_1/\widehat{p}_1, ..., \widehat{q}_J/\widehat{p}_J)$, son los siguientes (se resumen en el cuadro):

1º) Se busca el valor de i_1. Puede comprobarse que $M(\{1\}) = 0,1/0,3 = 1/3 < M(\{1,2\}) = (0,1 + 0,4)/(0,3 + 0,05) = 10/7$, y, por otro lado, $M(\{1,2\}) = 10/7 > M(\{1,2,3\}) = (0,1 + 0,4 + 0,38)/(0,3 + 0,05 + 0,5) = 88/85 > M(\{1,2,3,4\}) = 1$. Por lo tanto, $M(\{i_0 + 1, ..., i\})$ se maximiza con el conjunto $A = \{1,2\}$, y se tiene que $i_1 = 2$.

2º) Ahora se debe hallar el valor de $i_1 > i_0$. Para ello, calculamos $M(\{3\}$ y $M(\{3,4\})$. Vemos que $M(\{3\}) = 0,38/0,50 = 19/25 = 0,76 < M(\{3,4\}) = (0,38 + 0,12)/(0,5 + 0,15) = 10/13 = 0,7692$. Por consiguiente, $M(\{i_1 + 1, ..., i\})$ se maximiza con el conjunto $A = \{3,4\}$, y se tiene que $i_2 = 4$. Como $i_2 = J$, se da por terminado el proceso. Se han obtenido dos "level sets": $A_1 = \{1,2\}$ y $A_2 = \{3,4\}$. Con esto, se ha cubierto ya la fase 3 del algoritmo.

3º) Solo resta obtener las ponderaciones o proyección $E_{\widehat{p}}$. Para ello, basta asociar a cada entero j el valor de $M(A_k)$ para el conjunto A_k en el que se encuentra dicho entero. Obviamente, en este caso, a los enteros 1 y 2 les corresponde $M(A_1) = 10/7$, y a 3 y 4, $M(A_2) = 10/13$. Por lo tanto, la proyección es $E_{\widehat{p}} = (10/7, 10/7, 10/13, 10/13)$.

Finalmente, podemos ver cómo se construye ahora la estimación MV $\overline{p}$ bajo la hipótesis (5b). Dicha estimación resulta del producto elemento a elemento entre la proyección y la estimación MV $\widehat{p}$. Por lo tanto, se obtiene $\overline{p} = (0,3857, 0,0643, 0,05176, 0,15)$. Recordemos que la hipótesis (5b) exigía la condición $\sum_{j=1}^{i} p_j \geq \sum_{j=1}^{i} q_j$, para cualquier $i < J$ (también para $i = J$, pero esto se cumple por construcción) pero habiendo algún valor de i para el que la desigualdad sea estricta. Véase que $\widehat{p}$ no cumple dicha condición ni para $i = 2$ ni para $i = 3$, ya que $\widehat{p}_1 + \widehat{p}_2 = 0,35 < \widehat{q}_1 + \widehat{q}_2 = 0,50$ y $\widehat{p}_1 + \widehat{p}_2 + \widehat{p}_3 = 0,85 < \widehat{q}_1 + \widehat{q}_2 + \widehat{q}_3 = 0,88$. Pues bien, la proyección $E_{\widehat{p}}$ corrige $\widehat{p}$ de manera que se verifique (5b).[34] Puede comprobarse esto en las últimas columnas del cuadro a continuación.

[34] Por supuesto, hay infinitos vectores que verificarían la condición (5b). Pero de todos ellos, el de mayor verosimilitud es $\overline{p} = (\overline{p}_1, ..., \overline{p}_J)$, siendo $\overline{p}_i = E_{\widehat{p}}^{(i)} \widehat{p}_i$. Esta propiedad es demostrada por Robertson y Wright (1981), en su Teorema 2.1.

Ejemplo 1 de aplicación de LSA

i	$\widehat{p}_i$	$\widehat{q}_i$	$M(\{i_0+1,...,i\})$	$M(\{i_1+1,...,i\})$	$E^{(i)}_{\widehat{p}}$	$\overline{p}_i$	$\sum_{j=1}^{i}\widehat{p}_j$	$\sum_{j=1}^{i}\widehat{q}_j$	$\sum_{j=1}^{i}\overline{p}_j$
1	0,30	0,10	1/3	-	10/7	0,4286	0,30	0,10	0,4286
2	0,05	0,40	10/7	-	10/7	0,0714	0,35	0,50	0,50
3	0,50	0,38	88/85	19/25 = 0,76	10/13	0,3846	0,85	0,88	0,8846
4	0,15	0,12	1	10/13 = 0,7692	10/13	0,1154	1	1	1

Si $\widehat{p}$ hubiera cumplido la restricción (5b), la proyección $E_{\widehat{p}}$ resultaría ser un vector de unos, y, por tanto, se tendría que $\overline{p} = \widehat{p}$. Esto ocurre si modificamos convenientemente el vector $\widehat{p}$ del ejemplo anterior, para que sea ahora $\widehat{p} = (0,30, 0, 20, 0, 38, 0, 12)$. Al aplicar LSA, se tendrá un solo conjunto A ($A_1 = \{1, 2, 3, 4\}$), cuyo $M(A)$ asociado es $M(A_1) = 1$.[35] Véase el resultado de la implementación del algoritmo en el cuadro siguiente:

Ejemplo 2 de aplicación de LSA

i	$\widehat{p}_i$	$\widehat{q}_i$	$M(\{i_0+1,...,i\})$	$E^{(i)}_{\widehat{p}}$	$\overline{p}_i$	$\sum_{j=1}^{i}\widehat{p}_j$	$\sum_{j=1}^{i}\widehat{q}_j$	$\sum_{j=1}^{i}\overline{p}_j$
1	0,30	0,10	1/3	1	0,30	0,30	0,10	0,30
2	0,20	0,40	1	1	0,20	0,50	0,50	0,50
3	0,40	0,38	44/45	1	0,40	0,90	0,88	0,90
4	0,10	0,12	1	1	0,10	1	1	1

A.2.2. Cálculo de $P(l, J)$. Resultados de Barlow (1972)

La derivación de la distribución asintótica del estadístico de contraste de C3 requiere el cálculo de ciertas probabilidades $P(l, J)$. En realidad, dichas probabilidades se definen utilizando un tercer argumento, un vector de pesos w, que, en nuestro caso, es siempre el vector $\widehat{p}$, de dimensión J. La probabilidad $P(l, J; \widehat{p})$ representa la probabilidad de que la proyección $E_{\widehat{p}}$ contenga l valores distintos (análogamente, la probabilidad de que el algoritmo LSA expuesto arriba genere exactamente l "level sets"), dado un número de pérdidas J en la función de pérdidas discreta empleada (por tanto, el número máximo de level sets y de valores distintos en $E_{\widehat{p}}$ es J). La forma de calcular dichas probabilidades se expone en Barlow (1972), págs. 134-145. El proceso consiste en la aplicación recursiva de la ecuación (3.23) de Barlow (1972), utilizando además la condición $\sum_{l=1}^{J} P(l, J; w) = 1$ y la expresión analítica de las probabilidades $P(J, J; w)$, que ya ha sido obtenida en la literatura estadística para $J \leq 5$, y que luego expondremos. La ecuación citada es la siguiente:

$$P(l, J; w) = \sum_{\{B_1,...,B_l\}\in\zeta} P(l, l; (W(B_1), W(B_2), ..., W(B_l))') \left[\prod_{i=1}^{l} P(1, C(B_i); w(B_i))\right], \qquad (17)$$

siendo:

$B_1...B_l$ es una posible partición del conjunto $\{1, 2, ..., J\}$ formada por l subconjuntos, los cuales deben tener la característica de estar compuestos por enteros consecutivos. Por ejemplo, si $J = 4$ y $l = 3$, una posible partición del tipo especificado es $B_1 = \{1\}$, $B_2 = \{2, 3\}$, $B_3 = \{4\}$.[36]

ζ es el conjunto de todas esas posibles particiones de $\{1, 2, ..., J\}$ formadas por l subconjuntos de enteros consecutivos. Por ejemplo, en el caso $J = 4, l = 3$, $\zeta = \{\{1\}\{2, 3\}\{4\}; \{1, 2\}\{3\}\{4\}; \{1\}\{2\}\{3, 4\}\}$.

$W(B_i) = \sum_{k\in B_i} w_k$. Así, $(W(B_1), W(B_2)..., W(B_l))'$ constituye un vector de pesos de dimensión l.

$C(B_i)$ denota la cardinalidad del conjunto B_i.

[35] Recuérdese que si existen dos enteros $i \leq r$ que maximizan $M(\{1, ..., i\})$, se elige el mayor. Por eso en este caso, $i_1 = 4$, y solo hay un "level set" A.

[36] En el problema que nos ocupa (el correspondiente al test C3), la definición de los conjuntos B_i debe ser la que hemos expuesto (enteros consecutivos). Pero la ecuación (3.23) de Barlow (1972) aplica en otros contextos donde aparecen este tipo de probabilidades $P(l, J)$, en los que los conjuntos B_i no tienen porqué verificar dicha definición.

Finalmente, $w(B_i)$ es la parte del vector de pesos w correspondiente a los subíndices de B_i, y tendrá, por tanto, dimensión $C(B_i)$. Es decir, si $B_i = \{k_0, k_0+1, k_0+2\}$ (para algún entero k_0), $w(B_i) = (w_{k_0}, w_{k_0+1}, w_{k_0+2})$.

Por su parte, la expresión analítica de las probabilidades $P(J, J; w)$ es conocida para $J \leq 5$, y se presentan en las ecuaciones (3.17) y (3.19) de Barlow (1972), que reproducimos a continuación:

$$\begin{aligned} P(1,1;w) &= 1, \ \forall w \\ P(2,2;w) &= \frac{1}{2}, \ \forall w \\ P(3,3;w) &= \frac{1}{4} + \frac{1}{2\pi} arcsen\rho_{12} \\ P(4,4;w) &= \frac{1}{8} + \frac{1}{4\pi}(arcsen\rho_{12} + arcsen\rho_{23}) \\ P(5,5;w) &\approx \frac{1}{16} + \frac{1}{8\pi}(arcsen\rho_{12} + arcsen\rho_{23} + arcsen\rho_{34}) + \frac{1}{4\pi^2} arcsen\rho_{12} arcsen\rho_{34}, \end{aligned} \tag{18}$$

siendo

$$\rho_{i,i+1} = -\left(\frac{w_i w_{i+2}}{(w_i + w_{i+1})(w_{i+1} + w_{i+2})}\right)^{1/2}$$

Las cuatro primeras expresiones de (18) son exactas. Por su parte, $P(5,5;w)$ podría obtenerse de forma exacta usando las tablas de Abrahamson (1964), pero la aproximación de Plackett (1954) que presentamos en (18) funciona razonablemente bien.

Utilizando la ecuación (17), con las $P(l,l;w)$ dadas por (18), y teniendo en cuenta que $\sum_{l=1}^{J} P(l,J;w) = 1$, pueden obtenerse fácilmente las probabilidades $P(l,J;w)$, para cualquier $l \leq J$, siempre que $J \leq 5$. Para valores de $J > 5$, el cálculo de $P(l,J;w)$ se complicaría notablemente. Para dichos casos (o, en cualquier caso, si el usuario no desea asumir este coste de cálculo), aconsejamos usar la cota superior (13) para la implementación de C3. Vamos a mostrar un ejemplo de cómo se realizaría el cálculo de $P(l,J;w)$ en un caso sencillo, $P(2,3;w)$:

En este caso, solo hay dos particiones posibles: $B_1 = \{1\}, B_2 = \{2,3\}$ y $B_1' = \{1,2\}, B_2' = \{3\}$. Por lo tanto, aplicando (17):

$$\begin{aligned} P(2,3;w) = P(2,2;(w_1, w_2+w_3)')[P(1,1;w_1)P(1,2;(w_2,w_3)')] + \\ P(2,2;(w_1+w_2,w_3)')[P(1,2;(w_1,w_2)')P(1,1;w_3)] = \\ \frac{1}{2}[1P(1,2;(w_2,w_3)')] + \frac{1}{2}[P(1,2;(w_1,w_2)')1] = \frac{1}{2}\frac{1}{2} + \frac{1}{2}\frac{1}{2} = \frac{1}{2}, \ \forall w, \end{aligned}$$

donde se ha utilizado el valor de $P(2,2;w)$ en (18), mientras que $P(1,2;w)$ se deduce directamente de la restricción $P(1,2;w) + P(2,2;w) = 1$.

Barlow (1972) presenta muchos otros ejemplos de cálculo de probabilidades $P(l,J;w)$. Recuérdese que, en nuestro caso, $w = \widehat{p}$.

A.3. Aleatorización de tests discretos

Sea un test con estadístico de contraste λ, variable aleatoria discreta, cuya función de distribución es F_λ. Supongamos, por simplicidad, que la región de rechazo del test es la cola inferior, es decir, que, teóricamente, la forma de proceder del test consiste en que "se rechaza la hipótesis nula si y solo si $\lambda_0 \leq \lambda_c$", siendo λ_0 el valor observado de λ en la muestra, α el nivel de significación fijado o tamaño teórico del test, e λ_c el valor crítico, es decir, el valor del soporte de λ tal que $F(\lambda_c) = \alpha$. El problema es que, al ser λ una variable discreta y, por tanto, F_λ una función escalonada, salvo por casualidad, no existirá ningún valor λ_c en su soporte para el que dicha igualdad se verifique, y, por tanto, no es posible

construir una región crítica para el test exactamente con el tamaño deseado. Sean λ_b y λ_d los dos valores del soporte de λ más próximos a λ_c, por defecto y exceso, respectivamente. Es decir, son los dos valores contiguos del soporte que verifican $F_{(b)} = F_\lambda(\lambda_b) < \alpha < F_\lambda(\lambda_d) = F_{(d)}$. Entonces, existen dos opciones para resolver el test, elegir λ_b o λ_d como valores críticos, es decir:

– Rechazar si y solo si $\lambda_0 \leq \lambda_b$, en cuyo caso la probabilidad de rechazar la hipótesis nula con el test, o tamaño del contraste, es solo $F_{(b)} < \alpha$, de modo que el test será sesgado en tamaño, por defecto.

– Rechazar si y solo si $\lambda_0 \leq \lambda_d$, en cuyo caso la probabilidad de rechazar la hipótesis nula con el test, o tamaño del contraste, es $F_{(d)} > \alpha$, de modo que el test será sesgado en tamaño, por exceso.

Antes de seguir adelante, veamos un ejemplo numérico para reafirmar estas ideas:

Supongamos que $\lambda \sim B(10, 0{,}5)$ y $\alpha = 0,10$. Para los cuatro primeros puntos del soporte, la función de distribución vale: $F_\lambda(0) = 0,0010$, $F_\lambda(1) = 0,0107$, $F_\lambda(2) = 0,0547$ y $F_\lambda(3) = 0,1719$. Como vemos, no hay ningún punto del soporte para el que la función de distribución valga α. Los valores candidatos a utilizarse como puntos críticos son $\lambda_b = 2$ y $\lambda_d = 3$. De utilizarse el primero, el tamaño del test sería $F_{(b)} = F_\lambda(2) = 0,0547$, y si se emplea el segundo, el tamaño sería $F_{(d)} = F_\lambda(3) = 0,1719$, niveles lejanos del tamaño teórico fijado α.

Existen dos formas alternativas de abordar este problema:

a) La primera de ellas consiste simplemente en que el test actúe de forma automática cuando la decisión no es "conflictiva", y, en cambio, cuando lo sea, suministre la información que dispone para que sea el usuario quien tome la decisión que estime oportuna. Es decir, el test debería actuar del siguiente modo:

– Rechaza la hipótesis nula si $\lambda_0 \leq \lambda_b$.

– No rechaza la hipótesis nula si $\lambda_0 > \lambda_d$.

– Si $\lambda_0 = \lambda_d$ (ésta es la situación conflictiva), suministra al usuario los valores λ_b, λ_d, $F_{(b)}$ y $F_{(d)}$, para que éste tome la decisión. En realidad, eso supone que el usuario, sobre la marcha, redefina el riesgo de cometer errores de tipo I que desea asumir según sea el contexto o tipo de problema, amoldándose a los dos puntos más próximos a λ_c, es decir, elija entre no rechazar la hipótesis nula a nivel de significación $F_{(b)}$ o rechazarla a nivel de significación $F_{(d)}$.

b) La otra opción es diseñar *tests aleatorizados*. Se trata de usar un experimento aleatorio de Bernouilli con probabilidad adecuada p^* en los casos de duda ($\lambda_0 = \lambda_d$) para lograr que el tamaño del test sea, efectivamente, igual al deseado, α. La resolución del test sería la siguiente:

– Rechaza la hipótesis nula si $\lambda_0 \leq \lambda_b$.

– No rechaza la hipótesis nula si $\lambda_0 > \lambda_d$.

– Si $\lambda_0 = \lambda_d$, se lleva a cabo un experimento de Bernouilli, de probabilidad $p^* = \frac{\alpha - F_{(b)}}{F_{(d)} - F_{(b)}}$. Si el resultado del experimento es "éxito", se rechaza la hipótesis nula. Si es "fracaso", no se rechaza.

De este modo, sea X la variable aleatoria de Bernouilli, la probabilidad de rechazo o tamaño del test será igual a $P(\lambda \leq \lambda_b) + P((\lambda = \lambda_d) \wedge (X = 1)) = F_{(b)} + ((F_{(d)} - F_{(b)})p^*) = F_{(b)} + (\alpha - F_{(b)}) = \alpha$, tal y como se deseaba. El contraste tendrá el tamaño teórico, pero, en casos "conflictivos" ($\lambda_0 = \lambda_d$), se resuelve en última instancia en base al puro azar. Muchos estadísticos son reacios a este procedimiento, argumentando que dos usuarios podrían estar manejando la misma muestra y aplicando el mismo contraste y, sin embargo, llegando a conclusiones distintas sobre la hipótesis nula.

En la práctica, el uso del método a) ó b) depende de la situación en la que se encuentre el usuario del contraste. Si se está llevando a cabo éste de forma aislada (para una aplicación concreta), lo razonable es emplear la alternativa a). Si se está repitiendo sistemáticamente el test para un gran número de muestras, puede proceder b), al interesar que el test quede totalmente automatizado y garantizando, además, un tamaño correcto. Lógicamente, en nuestros ejercicios de simulación, aplicaremos el mecanismo de aleatorización para implementar los tests HM, B y C2.

Antes de dar por concluida la discusión sobre el asunto de la imposibilidad de construir regiones críticas de tamaño α en tests con distribución no continua, quedan dos comentarios relevantes que hacer: en primer lugar, hemos abordado todo el problema suponiendo que el contraste establece la región crítica en la cola inferior de la distribución. Pero todo el desarrollo anterior podría adaptarse de forma trivial a casos donde la cola de rechazo es la superior o ambas, sin ningún cambio conceptual. En segundo lugar, debemos hacer notar que el problema de la región crítica en tests con distribución no continua se atenúa conforme el número de elementos del soporte aumenta, ya que F_λ toma valores en $[0, 1]$, por lo que el número de "escalones" de F_λ aumentaría y su altura disminuiría. Es decir, al incrementarse el número de elementos del soporte de λ, F_λ se aproxima a una función continua, y el problema tiende a desaparecer. En el caso del

test Binomial, dicho número es exactamente el tamaño muestral más uno (porque el soporte del estadístico de contraste es el conjunto $\{0, 1, ..., T\}$). En el test HM, el soporte es el conjunto $\{0, 1, ..., n_1\}$, siendo n_1 un valor directamente relacionado con el tamaño muestral (su significado exacto se puede consultar en el apartado 2.1). Finalmente, el soporte del estadístico de contraste de C2 es más difícil de determinar, pero su cardinalidad depende positivamente de T y del número de pérdidas definido J (también depende de los valores concretos de las pérdidas $a_1, ..., a_J$). En general, podemos decir que si manejamos muestras relativamente grandes, el problema en cuestión tendrá muy poco efecto en cualquiera de los tres tests. En el caso de C2, también se mitiga conforme mayor es el número de pérdidas J (lo que, por el contrario, eleva el coste computacional del test).

B. Apéndice: Demostraciones de equivalencias entre tests

Se adjuntan aquí algunas demostraciones relacionadas con las equivalencia entre algunos tests, que quedaron pendientes cuando se realizó la exposición teórica de los mismos.

B.1. Equivalencia de hipótesis de los tests TC y P-T en el caso $m = 2$

Recordemos las hipótesis nulas de los dos tests:

$$\text{Hipótesis TC: } H_0^{TC} \equiv p_{ij} = p_i^y p_j^v, \text{ para } i = 1, ..., m,\ j = 1, ..., m$$
$$\text{Hipótesis P-T: } H_0^{PT} \equiv \sum_{i=1}^{m} p_{ii} = \sum_{i=1}^{m} p_i^y p_i^v,$$

donde p_{ij} denota la probabilidad del suceso $\{y_t \in r_i, v_t \in r_j\}$, p_i^y la probabilidad del suceso $\{y_t \in r_i\}$ y p_j^v la del suceso $\{v_t \in r_j\}$.

Obviamente, $H_0^{TC} \Rightarrow H_0^{PT}$ para cualquier valor de m. Se trata de demostrar que en el caso $m = 2$, además, también ocurre $H_0^{PT} \Rightarrow H_0^{TC}$, y, por tanto, las hipótesis de ambos tests son la misma. Por consiguiente, la afirmación a probar es:

$$p_{11} + p_{22} = p_1^y p_1^v + p_2^y p_2^v \Rightarrow p_{11} = p_1^y p_1^v,\ p_{12} = p_1^y p_2^v,\ p_{21} = p_2^y p_1^v,\ p_{22} = p_2^y p_2^v. \tag{19}$$

Utilizaremos las definiciones de las probabilidades marginales para el desarrollo de la prueba, a saber:

$$p_1^y = p_{11} + p_{12},\ p_2^y = p_{21} + p_{22},\ p_1^v = p_{11} + p_{21},\ p_2^v = p_{12} + p_{22}. \tag{20}$$

Ya estamos en condiciones de demostrar (19):

1. Demostración de $p_{11} + p_{22} = p_1^y p_1^v + p_2^y p_2^v \Rightarrow p_{22} = p_2^y p_2^v$:

$p_{11} + p_{22} = p_1^y p_1^v + p_2^y p_2^v = (1 - p_2^y)(1 - p_2^v) + p_2^y p_2^v = 1 - p_2^v - p_2^y + 2p_2^y p_2^v = 1 - (p_{12} + p_{22}) - (p_{21} + p_{22}) + 2p_2^y p_2^v = 1 - (1 - p_{11} + p_{22}) + 2p_2^y p_2^v$, habiéndose utilizado en un paso intermedio las definiciones en (20). Por lo tanto, $p_{11} + p_{22} = p_{11} - p_{22} + 2p_2^y p_2^v$. Simplificando, se tiene: $p_{22} = p_2^y p_2^v$, q.e.d.

2. Demostración de $p_{11} + p_{22} = p_1^y p_1^v + p_2^y p_2^v \Rightarrow p_{11} = p_1^y p_1^v$:

Se deduce inmediatamente de la parte izquierda de la implicación y del resultado 1.

3. Demostración de $p_{11} + p_{22} = p_1^y p_1^v + p_2^y p_2^v \Rightarrow p_{12} = p_1^y p_2^v$:

Multiplicando por -1 y sumando 1 en $p_{11} + p_{22} = p_1^y p_1^v + p_2^y p_2^v$, se tiene que $1 - p_{11} + p_{22} = p_{12} + p_{21} = 1 - p_1^y p_1^v - p_2^y p_2^v = 1 - p_1^y(1 - p_2^v) - (1 - p_1^y)p_2^v = 1 - p_1^y + p_1^y p_2^v - p_2^v + p_1^y p_2^v = 1 - p_1^y - p_2^v + 2p_1^y p_2^v = 1 - (p_{11} + p_{12}) - (p_{12} + p_{22}) + 2p_1^y p_2^v = 1 - (1 - p_{21} + p_{12}) + 2p_1^y p_2^v$, habiéndose utilizado en un paso intermedio las definiciones en (20). Por lo tanto, $p_{12} + p_{21} = p_{21} - p_{12} + 2p_1^y p_2^v$. Simplificando, se tiene: $p_{12} = p_1^y p_2^v$, q.e.d.

4. Demostración de $p_{11} + p_{22} = p_1^y p_1^v + p_2^y p_2^v \Rightarrow p_{21} = p_2^y p_1^v$:

Dado que $1 - p_{21} = p_{11} + p_{12} + p_{22}$, aplicando ahora los resultandos 1, 2 y 3, se tiene que $1 - p_{21} = p_1^y p_1^v + p_1^y p_2^v + p_2^y p_2^v = p_1^y(p_1^v + p_2^v) + p_2^y p_2^v = p_1^y + p_2^y p_2^v$, donde se ha utilizado que $p_1^v + p_2^v = 1$, por definición. Por lo tanto, $1 - p_{21} = p_1^y + p_2^y p_2^v = (1 - p_2^y) + p_2^y p_2^v = 1 - p_2^y(1 - p_2^v) = 1 - p_2^y p_1^v$. Simplificando, $p_{21} = p_2^y p_1^v$, q.e.d.

Por consiguiente, queda demostrada la afirmación (19)), es decir, queda demostrada la equivalencia de hipótesis nulas de los tests TC y P-T en el caso $m = 2$.

B.2. Equivalencias entre los tests P-T y C1-v2 en caso de función discreta "simétrica"

En la presentación del test C1-v2 (sección 4.2) se afirmó que, en el caso de utilizar una función de pérdida discreta "simétrica" para la implementación de C1-v2, este test y P-T coincidían en muchos aspectos de su especificación, aunque diferían en la hipótesis alternativa (por lo que, ni siquiera en esta

situación, los dos tests son equivalentes). Decimos que la función de pérdida es "simétrica" si solamente asigna dos valores: uno, a_1, a los cuadrantes (i,i), y el otro, a_2, al resto, siendo $a_1 < a_2$. En tal situación, la coincidencia entre los tests C1-v2 y P-T se produce en cuanto a sus hipótesis nulas y al valor absoluto de sus estadísticos de contraste. Éstas son las dos afirmaciones que se demostrarán en este apartado, siendo su expresión formal:

Sea la función de pérdida usada en C1-v2 "simétrica", entonces:

$$H_0^{C1v2} \equiv ap = aq \Leftrightarrow H_0^{PT} \equiv \sum_{i=1}^{m} p_{ii} = \sum_{i=1}^{m} p_i^y p_i^v, \text{ y} \tag{21}$$

$$C_{1,2} = -S_{PT} \tag{22}$$

donde $C_{1,2}$ y S_{PT} denotan los estadísticos de contraste de los tests C1-v2 y P-T, respectivamente. Las definiciones de p_{ij}, p_i^y y p_j^v fueron recordadas al principio del apartado B.1. Por su parte, p y q representan vectores $J \times 1$ $p = (p_1, ..., p_J)'$, $q = (q_1, ..., q_J)'$, tales que p_i la probabilidad de que (y_t, v_t) tenga asociada una pérdida de valor a_i, mientras q_i es la misma probabilidad pero bajo el supuesto de independencia estocástica entre datos y previsiones.

Recuérdese también que p_i y q_i se construyen por $p_i = \sum_{(r_k,r_q)\in R_i} p_{kq}$ y $q_i = \sum_{(r_k,r_q)\in R_i} p_k^y p_q^v$, respectivamente, siendo R_i el conjunto de cuadrantes (r_k, r_q) a los que se asignó pérdida a_i.

1. Demostración de la proposición (21):

En el caso de función de pérdida "simétrica" (recuérdese que solo hay dos pérdidas: a_1 (en la diagonal) y a_2 (fuera de ésta)), se tiene que:

$ap = a_1 \sum_{i=1}^{m} p_{ii} + a_2 \sum_{i=1}^{m} \sum_{j=1, j\neq i}^{m} p_{ij} = a_1 \sum_{i=1}^{m} p_{ii} + a_2(1 - \sum_{i=1}^{m} p_{ii}) = a_2 + (a_1 - a_2) \sum_{i=1}^{m} p_{ii}$, ya que $\sum_{i=1}^{m} \sum_{j=1}^{m} p_{ij} = 1$.

Análogamente, $aq = a_2 + (a_1 - a_2) \sum_{i=1}^{m} p_i^y p_i^v$.

Por lo tanto, $ap = aq \Rightarrow a_2 + (a_1 - a_2) \sum_{i=1}^{m} p_{ii} = a_2 + (a_1 - a_2) \sum_{i=1}^{m} p_i^y p_i^v \Rightarrow \sum_{i=1}^{m} p_{ii} = \sum_{i=1}^{m} p_i^y p_i^v$. Se ha demostrado que $H_0^{C1v2} \Rightarrow H_0^{PT}$.

Reconstruyendo los pasos en el sentido contrario obtenemos la otra implicación, $H_0^{PT} \Rightarrow H_0^{C1v2}$:

$\sum_{i=1}^{m} p_{ii} = \sum_{i=1}^{m} p_i^y p_i^v \Rightarrow a_2 + (a_1 - a_2) \sum_{i=1}^{m} p_{ii} = a_2 + (a_1 - a_2) \sum_{i=1}^{m} p_i^y p_i^v \Rightarrow a_1 \sum_{i=1}^{m} p_{ii} + a_2(1 - \sum_{i=1}^{m} p_{ii}) = a_1 \sum_{i=1}^{m} p_i^y p_i^v + a_2(1 - \sum_{i=1}^{m} p_i^y p_i^v) \Rightarrow$

$a_1 \sum_{i=1}^{m} p_{ii} + a_2 \sum_{i=1}^{m} \sum_{j=1, j\neq i}^{m} p_{ij} = a_1 \sum_{i=1}^{m} p_i^y p_i^v + a_2 \sum_{i=1}^{m} \sum_{j=1, j\neq i}^{m} p_{ij} \Rightarrow ap = aq$, por definición de función de pérdida "simétrica". Por tanto, también se ha demostrado $H_0^{PT} \Rightarrow H_0^{C1v2}$, con lo que la proposición (21) queda probada.

2. Demostración de la proposición (22):

Recuérdense primero las expresiones de los estadísticos de contraste S_{PT} y E_2:

$S_{PT} = \sqrt{T} h(\widehat{P}) \widehat{W}^{-1/2}$, siendo $h(P) = \sum_{i=1}^{m} p_{ii} - \sum_{i=1}^{m} p_i^y p_i^v$ y $W = \left(\frac{\partial h}{\partial P}\right) V_P \left(\frac{\partial h}{\partial P}\right)'$, y V_P cierta matriz simétrica $m^2 \times m^2$. El estimador de W es, simplemente, $\widehat{W} = W_{P=\widehat{P}}$.

$C_{1,2} = \sqrt{T} a f(\widehat{P}) \widehat{G}_p^{-1/2}$, siendo $f(P) = p - q$ y $G_p = a \nabla f(P) V_P \nabla f(P)' a'$. El estimador de G_p es, simplemente, $\widehat{G}_p = [G_p]_{P=\widehat{P}}$.

Pues bien, en el caso de que la función de pérdida empleada en C1-v2 sea "simétrica", se tiene:

$$1)\ f(P) = p - q = \begin{pmatrix} \sum_{i=1}^{m} p_{ii} - \sum_{i=1}^{m} q_{ii} \\ \sum_{i=1}^{m} \sum_{j=1, j\neq i}^{m} p_{ij} - \sum_{i=1}^{m} \sum_{j=1, j\neq i}^{m} q_{ij} \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{m} p_{ii} - \sum_{i=1}^{m} p_i^y p_i^v \\ (1 - \sum_{i=1}^{m} p_{ii}) - (1 - \sum_{i=1}^{m} p_i^y p_i^v) \end{pmatrix} = \begin{pmatrix} h(P) \\ -h(P) \end{pmatrix} =$$

$\begin{pmatrix} 1 \\ -1 \end{pmatrix} h(P)$, siendo $h(P)$ la función definida para el estadístico S_{PT}.

2) $\nabla f(P) = \begin{pmatrix} 1 \\ -1 \end{pmatrix} \left(\frac{\partial h}{\partial P}\right)$, siendo $\left(\frac{\partial h}{\partial P}\right)$ el vector gradiente $1 \times m^2$ definido para S_{PT}.

3) $G_p = \begin{pmatrix} a_1 & a_2 \end{pmatrix} \begin{pmatrix} 1 \\ -1 \end{pmatrix} \left(\frac{\partial h}{\partial P}\right) V_P \left(\frac{\partial h}{\partial P}\right) \begin{pmatrix} 1 & -1 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = W \left(a_1 - a_2\right)^2$, siendo W la varianza (escalar) definida para S_{PT}.

4) Finalmente, con los resultados 1 a 3, se tiene que $E_2 = \sqrt{T}\widehat{W}^{-1/2} \left|a_1 - a_2\right|^{-1} \begin{pmatrix} a_1 & a_2 \end{pmatrix} \begin{pmatrix} 1 \\ -1 \end{pmatrix} h(\widehat{P}) = \sqrt{T}h(\widehat{P})\widehat{W}^{-1/2} \frac{a_1 - a_2}{|a_1 - a_2|}$. Dado que las pérdidas verifican $a_1 < a_2$, se obtiene que $C_{1,2} = -\sqrt{T}h(\widehat{P})\widehat{W}^{-1/2} = -S_{PT}$, q.e.d. Fijémonos que, en este caso especial, el valor de las pérdidas es irrelevante para C1-v2.

B.3. Relación entre la hipótesis alternativa de C3 y la de C1-v1, C1-v2 y C2

Se pretende demostrar que la hipótesis alternativa de C3, denotada por H_1', es condición suficiente para la hipótesis del resto de tests que proponemos en el capítulo, denotada por H_1. Es decir, la proposición a demostrar es $H_1' \Rightarrow H_1$. En cambio, como ya se comentó en el apartado 3.3, la implicación de sentido contrario no es cierta. Recordamos las hipótesis en cuestión:

$$\begin{gathered} H_1 \equiv ap < aq \\ H_1' \equiv H_2 - H_0', \text{ siendo:} \\ H_0' \equiv p = q, \\ H_2 \equiv \sum_{j=1}^{i} p_j \geq \sum_{j=1}^{i} q_j, \ i = 1, ..., J, \end{gathered}$$

siendo $(a_1, a_2, ..., a_J)$ el vector fila con las J pérdidas distintas de la función de pérdidas discreta, verificando $a_1 < a_2 < ... < a_J$, mientras p_i representa la probabilidad teórica de obtener la pérdida a_i, y q_i es la probabilidad teórica de obtener dicha pérdida si datos y previsiones fueran independientes.

El razonamiento para la prueba es muy sencillo:

1. Sea un vector $p = (p_1, ..., p_J)$ dado, comprobaremos que el vector $q = (q_1, ..., q_J)$ para el que la pérdida asociada aq es mínima, de entre todos aquellos que verifican H_2, es el vector $q^* = p$.

2. Una vez comprobado esto, se tiene que todos los vectores q verificando H_1' (que excluye precisamente el caso $q = p$), cumplen la condición $aq > ap$. En consecuencia, se habrá demostrado que $H_1' \Rightarrow H_1$.

Solo necesitamos demostrar 1. Los pasos son los siguientes:

a) Primero, comprobemos que otra forma de escribir H_2 es: $H_2 \equiv \sum_{j=s}^{J} p_j \leq \sum_{j=s}^{J} q_j$, $s = J, ..., 1$. Esto es obvio, basta multiplicar por -1, y sumar 1 en ambos lados de la desigualdad original de H_2: $1 - \sum_{j=1}^{i} p_j = \sum_{j=i+1}^{J} p_j \leq 1 - \sum_{j=1}^{i} q_j = \sum_{j=i+1}^{J} q_j$, $i = 1, ..., J-1$. Por lo tanto, la desigualdad $\sum_{j=s}^{J} p_j \leq \sum_{j=s}^{J} q_j$ se verifica para $s = 2, ..., J$. Además, en el caso $s = 1$, se tiene que $\sum_{j=s}^{J} p_j = 1 = \sum_{j=s}^{J} q_j$. Entonces, la expresión H_2 puede escribirse como se proponía.

b) Si se pretende minimizar aq, la estrategia óptima es elegir el menor valor posible para q_J de entre los que verifican H_2, ya que $a_J > a_{J-1} > ... > a_1$. Como H_2 exige $p_J \leq q_J$, debe elegirse $q_J^* = p_J$.

c) Una vez tomada esta decisión, la estrategia óptima para minimizar aq, es elegir el menor valor posible para q_{J-1} de entre los que verifican H_2. Como H_2 exige $p_J + p_{J-1} \leq q_J + q_{J-1}$ y ya se eligió $q_J^* = p_J$, habrá que escoger $q_{J-1}^* = p_{J-1}$.

d) Razonando sucesivamente de este modo, se tiene que el vector q con menor pérdida aq de entre los que verifican H_2 es $q^* = p$, q.e.d.

C. Apéndice: Autocorrelaciones muestrales de las pérdidas

Se presentan aquí las tablas que dan respaldo a nuestra decisión de no aplicar el procedimiento de Bonferroni en las simulaciones correspondientes a los Experimentos 3 y 4 de la sección 5. Léanse los comentarios al respecto en el apartado 5.1.2. En primer lugar, en las Tablas 5 y 6 se exponen los promedios de las estimaciones del coeficiente de autocorrelación muestral de orden uno para las pérdidas $z_1, ..., z_T$, estimación que se denota por $\widehat{\gamma}_z(1)$. Por otro lado, para la muestra de pérdidas correspondiente a cada realización del experimento, se lleva a cabo un test para contrastar $\gamma_z(1) = 0$ contra $\gamma_z(1) \neq 0$, y los resultados se presentan en las Tablas 7 y 8. Dicho test se basa en la distribución asintótica de $\widehat{\gamma}_z(1)$, distribución que, en base a las fórmulas de Bartlett para la varianza de los coeficientes de autocorrelación, es $\widehat{\gamma}_z(1) \overset{a}{\sim} N(\gamma_z(1), \frac{T-1}{T(T+2)})$,[37]. El nivel de significación empleado en el test es $\alpha = 0,05$.

Tabla 6. Autocorr. orden 1 Pérdidas z_t
Experimento 1

T	θ					
	0,204	0,314	0,436	0,577	0,750	0,927
10	-0,084	-0,093	-0,108	-0,132	-0,125	-0,147
25	-0,035	-0,038	-0,045	-0,072	-0,075	-0,114
50	-0,005	-0,010	-0,015	-0,046	-0,053	-0,092
100	0,004	0,005	-0,004	-0,021	-0,036	-0,072

Promedio realizaciones $\widehat{\gamma}_z(1)$ en el experimento

Tabla 7. Autocorr. orden 1 Pérdidas z_t
Experimento 2

T	θ			
	0,209	0,333	0,500	0,905
10	-0,085	-0,103	-0,104	-0,136
25	-0,034	-0,045	-0,057	-0,085
50	-0,011	-0,019	-0,045	-0,066
100	-0,001	-0,010	-0,037	-0,056

Promedio realizaciones $\widehat{\gamma}_z(1)$ en el experimento

Tabla 8. Frecuencia No Rechazo de
$H_0 : \gamma_z(1) = 0$. Experimento 1

T	θ					
	0,204	0,314	0,436	0,577	0,750	0,927
10	0,98	0,99	0,97	0,98	0,97	0,97
25	0,96	0,96	0,95	0,94	0,93	0,93
50	0,96	0,96	0,95	0,94	0,91	0,91
100	0,96	0,95	0,94	0,93	0,90	0,88

Distribución del contraste: $\widehat{\gamma}_z(1) \overset{a}{\sim} N(0, \frac{T-1}{T(T+2)})$.

Tabla 9. Frecuencia No Rechazo de
$H_0 : \gamma_z(1) = 0$. Experimento 2

T	θ			
	0,209	0,333	0,500	0,905
10	0,99	0,98	0,98	0,97
25	0,97	0,96	0,97	0,95
50	0,96	0,96	0,96	0,94
100	0,96	0,95	0,95	0,92

Distribución de contraste: $\widehat{\gamma}_z(1) \overset{a}{\sim} N(0, \frac{T-1}{T(T+2)})$.

[37] Se denota por T la longitud de la muestra utilizada en el cálculo de $\widehat{\gamma}_z(1)$.

D. Apéndice: Estimación de probabilidades de rechazo de los contrastes bajo partición en 4 regiones

Se adjuntan a continuación las Tablas con las probabilidades de rechazo estimadas en las simulaciones para los Experimentos 1 a 4 de la sección 5, pero usando una partición de 4 regiones y función de pérdidas (16), en vez de la partición de 3 regiones con función de pérdidas (15) usada para obtener los resultados presentados entonces:

Tabla 10. Probabilidad Rechazo Experimento 1. Partición 4 regiones.
$y_t = \varepsilon_t - \theta\varepsilon_{t-1},\ \varepsilon_t \overset{iid}{\sim} N(0,0{,}1)$; $v_t = -\widehat{\theta}\widehat{\varepsilon}_{t-1}$; $\alpha = 5\,\%$, 1000 repeticiones

θ	ρ_{yv}	T	B	H-M	TC 2×2	TC $m\times m$	P-T	C1-v1	C1-v2	C2	C3
0,204	0,2	10	9,5	11,2	10,9	0,1	9,7	11,3	19,9	8,1	3,4
		25	15,8	19,3	13,8	1,6	8,5	16,1	21,7	14,4	6,8
		50	23,0	25,6	18,9	4,0	6,6	21,4	23,7	19,7	13,9
		100	34,3	39,0	27,7	9,2	6,9	30,2	32,6	29,2	23,3
0,314	0,3	10	13,3	17,3	14,2	0,8	14,4	20,3	26,6	15,9	9,4
		25	23,2	27,0	20,9	6,5	12,1	32,0	33,7	28,2	20,1
		50	35,6	38,0	29,8	13,4	13,7	44,8	44,7	41,5	30,3
		100	60,7	64,3	54,2	31,9	22,6	67,5	66,8	65,7	53,1
0,436	0,4	10	18,8	19,5	18,0	0,9	16,5	30,5	37,2	23,0	16,6
		25	35,6	36,9	28,6	9,7	17,1	50,5	51,2	45,6	31,2
		50	57,2	59,8	52,5	29,3	32,3	74,2	73,7	71,5	55,0
		100	84,2	86,4	79,9	67,7	58,6	94,1	93,9	93,6	85,5
0,577	0,5	10	27,8	26,5	25,2	1,9	23,7	42,7	48,2	33,6	26,3
		25	50,1	51,6	41,6	19,3	33,3	71,0	71,1	67,0	47,5
		50	77,1	78,8	72,5	51,4	55,8	92,1	91,7	90,6	76,6
		100	97,0	97,9	95,5	91,6	87,1	99,7	99,7	99,7	91,3
0,750	0,6	10	32,2	30,0	30,8	1,8	29,8	53,4	61,0	42,8	31,3
		25	64,0	65,5	57,0	32,6	50,8	83,3	84,9	81,7	63,4
		50	90,6	92,0	88,9	80,4	83,1	98,0	98,2	97,8	92,0
		100	99,3	99,6	99,3	99,3	97,8	100,0	100,0	100,0	99,7
0,927	0,68	10	44,1	39,0	40,4	3,3	36,3	66,0	70,2	55,1	44,0
		25	79,9	78,5	71,6	45,2	67,5	95,3	94,7	93,9	82,2
		50	98,4	98,4	97,0	92,7	94,0	100,0	100,0	100,0	99,0
		100	100,0	100,0	100,0	100,0	99,9	100,0	100,0	100,0	100,0

Tabla 11. Probabilidad Rechazo Experimento 2. Partición 4 regiones

$y_t = \varepsilon_t - \theta\varepsilon_{t-1},\ \varepsilon_t \overset{iid}{\sim} N(0,0{,}1)$; $v_t = -\widehat{\phi} y_{t-1}$; $\alpha = 5\,\%$, 1000 repeticiones

θ	ρ_{yv}	T	B	H-M	TC 2×2	TC $m\times m$	P-T	C1-v1	C1-v2	C2	C3
$0,209$	$0,2$	10	10,0	12,0	11,4	0,1	9,9	12,2	19,8	8,5	3,0
		25	15,2	17,2	12,0	2,1	8,0	15,4	19,2	13,4	8,2
		50	21,2	23,7	17,7	4,0	6,4	19,8	22,8	18,7	13,4
		100	30,9	35,9	26,3	8,2	6,6	30,7	32,8	29,9	21,6
$0,333$	$0,3$	10	15,5	16,3	17,2	1,7	12,4	21,9	27,5	17,5	11,8
		25	24,6	26,5	18,2	7,1	11,0	31,2	32,1	27,6	18,8
		50	40,1	43,3	33,8	13,3	14,9	47,7	47,2	44,9	31,6
		100	61,0	66,4	54,2	32,9	23,4	69,3	69,4	67,6	54,3
$0,500$	$0,4$	10	19,6	20,5	22,1	1,4	16,3	32,9	37,0	24,8	17,2
		25	35,9	36,7	28,7	11,3	19,1	50,6	49,8	44,7	30,6
		50	60,9	64,0	54,6	30,6	31,3	73,1	71,7	69,6	52,7
		100	87,6	89,0	81,4	65,6	54,5	93,5	93,0	92,7	84,8
$0,905$	$0,49$	10	24,7	25,7	27,2	2,7	20,8	41,6	44,5	32,1	24,2
		25	51,1	53,0	42,3	18,6	31,3	68,2	67,4	64,0	46,2
		50	78,9	81,0	74,2	48,4	53,8	90,1	89,3	88,7	75,2
		100	96,7	97,1	95,4	90,5	85,1	99,4	99,3	99,3	97,0

Tabla 12. Probabilidad Rechazo Experimento 3. Partición 4 regiones

$y_t = \beta_1 x_{1t} + \beta_2 x_{2t} + \varepsilon_t$, $(x_{1t}, x_{2t}, \varepsilon_t)' \overset{iid}{\sim} N(0_{3\times1}, \Sigma)$, $\Sigma = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \sigma_2^2 & 0 \\ 0 & 0 & 0,1 \end{pmatrix}$; $v_t = \widehat{\beta}_1 x_{1t}$,

$\alpha = 5\,\%$, 1000 repeticiones

σ_2^2	ρ_{yv}	T	B	H-M	TC 2×2	TC $m\times m$	P-T	C1-v1	C1-v2	C2	C3
23,9	0,2	10	9,8	8,6	10,3	0,0	9,0	10,5	14,6	7,4	2,7
		25	14,1	12,8	10,4	2,0	7,9	13,4	15,2	11,9	6,8
		50	20,4	21,0	14,9	4,2	6,6	19,5	22,0	18,6	11,5
		100	33,1	33,7	23,4	8,3	8,4	28,5	30,8	27,0	21,0
10,01	0,3	10	14,0	11,5	11,6	0,7	12,3	18,0	20,3	12,9	7,1
		25	23,9	21,0	16,2	4,5	9,7	27,5	28,1	23,3	16,7
		50	35,1	35,1	27,5	13,7	12,6	40,8	40,9	38,7	28,9
		100	58,3	59,4	50,5	35,2	22,9	65,1	64,7	63,8	51,4
2,9	0,5	10	27,2	22,1	22,9	3,1	22,3	36,0	39,2	29,8	23,8
		25	50,7	47,4	39,1	19,2	32,1	63,4	63,4	59,8	42,5
		50	76,8	75,2	68,7	48,4	54,9	86,6	86,1	85,2	73,7
		100	96,2	96,3	92,3	88,6	83,4	99,3	99,3	99,3	96,3
0,94	0,7	10	46,1	36,1	36,9	7,3	42,3	61,7	62,7	53,2	41,5
		25	81,9	76,6	71,2	53,7	70,2	92,2	91,5	90,1	78,9
		50	97,6	97,2	95,1	92,5	94,2	99,6	99,6	99,4	98,2
		100	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
0,134	0,9	10	81,5	70,6	73,3	25,1	77,9	93,4	92,4	88,9	80,5
		25	99,2	98,6	98,0	97,7	98,8	99,9	99,9	99,9	99,7
		50	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
		100	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
0,008	0,95	10	91,6	79,0	83,4	39,7	89,3	98,0	97,0	95,8	92,5
		25	99,9	99,9	99,9	99,9	99,8	100,0	100,0	100,0	100,0
		50	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
		100	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0

Tabla 13. Probabilidad Rechazo Experimento 4. Partición 4 regiones
$(y_t, v_t) \sim N(0_{2\times1}, \Sigma)$, siendo $\Sigma = \begin{pmatrix} \sigma_y^2 & \rho\sigma_y\sigma_v \\ \rho\sigma_y\sigma_v & \sigma_v^2 \end{pmatrix}$, $\sigma_y^2 = \sigma_v^2 = 1$,
$\alpha = 5\,\%$, 1000 repeticiones

$\rho_{yv} = \rho$	T	B	H-M	TC 2×2	TC $m\times m$	P-T	C1-v1	C1-v2	C2	C3
0,0	10	3,5	4,3	5,6	1,3	10,9	5,9	9,2	4,0	1,9
	25	4,7	4,4	6,6	4,5	9,6	5,6	6,1	4,9	2,9
	50	4,9	5,3	7,9	5,2	6,4	6,0	6,1	5,8	3,1
	100	5,3	5,3	5,5	4,7	5,4	6,1	6,3	5,9	3,2
0,2	10	11,5	9,5	9,3	1,9	12,4	14,1	17,2	10,1	6,6
	25	15,5	15,4	10,6	5,7	11,7	18,9	20,2	17,8	10,9
	50	19,9	19,9	13,9	9,7	12,6	23,8	25,1	23,0	12,4
	100	34,8	36,4	25,3	14,0	21,7	42,5	43,6	41,8	24,7
0,3	10	15,2	12,0	13,1	1,6	14,6	19,9	22,7	14,3	8,7
	25	24,2	23,8	17,1	9,1	18,0	31,1	33,3	28,9	15,4
	50	39,0	38,2	30,6	16,2	27,9	47,1	48,9	46,4	28,8
	100	59,7	59,5	50,0	34,0	48,8	73,8	74,5	73,4	49,8
0,4	10	22,3	18,8	18,5	2,3	19,1	27,3	30,5	21,6	13,4
	25	37,5	35,0	26,8	14,5	26,9	44,9	47,1	42,6	24,5
	50	57,1	55,9	47,4	31,6	47,3	69,2	70,1	67,9	47,1
	100	81,8	82,6	75,9	60,3	75,6	91,5	92,0	90,8	77,6
0,5	10	25,4	21,9	19,8	3,4	26,4	34,9	39,2	28,0	18,8
	25	49,2	46,6	39,0	23,2	41,0	62,9	63,8	58,7	39,6
	50	76,3	74,6	67,5	49,2	69,3	86,6	87,0	85,7	68,6
	100	95,8	95,5	91,9	86,2	94,4	98,8	98,8	98,6	95,6
0,7	10	50,6	40,4	42,3	5,8	43,9	62,0	65,0	53,8	39,9
	25	83,8	81,4	76,2	55,3	77,2	90,8	92,0	89,4	78,3
	50	98,7	97,6	96,7	91,9	97,4	99,7	99,7	99,5	98,4
	100	100,0	100,0	99,9	100,0	100,0	100,0	100,0	100,0	100,0
0,9	10	81,6	67,4	71,8	24,3	77,3	92,6	91,6	87,9	79,5
	25	99,0	98,6	97,7	96,3	98,4	99,9	99,8	99,7	99,5
	50	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	100	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
0,95	10	90,2	81,0	83,4	37,6	89,0	96,7	96,4	95,1	90,4
	25	100,0	99,8	99,9	100,0	100,0	100,0	100,0	100,0	100,0
	50	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	100	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
−0,25	10	1,0	1,8	9,5	1,8	17,7	1,4	2,9	0,7	0,9
	25	0,3	0,6	11,4	6,3	16,3	0,6	0,6	0,6	0,7
	50	0,3	0,2	21,8	11,7	21,4	0,2	0,2	0,2	0,3
	100	0,0	0,1	39,0	21,7	37,0	0,0	0,0	0,0	0,2
−0,8	10	0,0	0,0	52,0	13,5	43,4	0,0	0,1	0,0	0,1
	25	0,0	0,0	87,5	80,2	66,5	0,0	0,0	0,0	0,3
	50	0,0	0,0	99,4	98,7	88,8	0,0	0,0	0,0	0,3
	100	0,0	0,0	100,0	100,0	99,7	0,0	0,0	0,0	0,1

E. Apéndice: Pérdidas medias muestrales de los Experimentos

Se adjuntan a continuación las pérdidas observadas y las obtenidas, a partir de éstas, bajo el supuesto de independencia estocástica entre datos y previsiones. Es decir, son las estimaciones de los parámetros ap y aq involucrados en las hipótesis de los contrastes C1-v1, C1-v2 y C2. Las estimaciones se obtuvieron promediando los 1000 valores asociados a las estimaciones muestrales $a\widehat{p}$ y $a\widehat{q}$ que se generaron en las correspondientes repeticiones de los experimentos. Se adjuntan solo los resultados para longitud muestral $T = 100$.

Tabla 14. Pérdidas Medias. Experimentos 1 a 3. Longitud Muestral: $T = 100$

		Partición de 3 regiones		Partición de 4 regiones	
		Observ. $(a\widehat{p})$	Indep. $(a\widehat{q})$	Observ. $(a\widehat{p})$	Indep. $(a\widehat{q})$
θ	ρ_{yv}	Experimento 1			
0,204	0,2	0,695	0,715	1,275	1,360
0,314	0,3	0,694	0,776	1,215	1,402
0,436	0,4	0,678	0,836	1,151	1,443
0,577	0,5	0,647	0,889	1,062	1,476
0,750	0,6	0,607	0,930	0,981	1,505
0,927	0,68	0,543	0,930	0,884	1,506
θ	ρ_{yv}	Experimento 2			
0,209	0,2	0,694	0,712	1,273	1,361
0,333	0,3	0,696	0,775	1,217	1,400
0,500	0,4	0,675	0,837	1,152	1,442
0,905	0,49	0,643	0,886	1,060	1,476
σ_2^2	ρ_{yv}	Experimento 3			
23,9	0,2	0,692	0,711	1,272	1,360
10,01	0,3	0,697	0,773	1,218	1,401
2,9	0,5	0,647	0,887	1,061	1,477
0,94	0,7	0,525	0,955	0,837	1,522
0,134	0,9	0,322	0,995	0,515	1,548
0,008	0,95	0,229	1,003	0,373	1,554

Promedios medias muestrales pérdidas: observadas y bajo independencia datos-prevs

Tabla 15. Pérdidas Medias. Experimento 4. Longitud Muestral: $T = 100$

	Partición de 3 regiones		Partición de 4 regiones	
ρ_{yv}	Observ. (ap)	Indep. (aq_0)	Observ. (ap)	Indep. (aq_0)
0,0	1,008	1,011	1,552	1,559
0,2	0,897	1,009	1,372	1,559
0,3	0,839	1,009	1,283	1,559
0,4	0,784	1,011	1,185	1,559
0,5	0,712	1,011	1,087	1,559
0,7	0,557	1,010	0,852	1,559
0,9	0,330	1,012	0,516	1,559
0,95	0,231	1,010	0,373	1,558
-0,25	1,145	1,010	1,777	1,558
-0,8	1,409	1,011	2,313	1,559

Promedios medias muestrales pérdidas: observadas y bajo independencia datos-prevs

F. Apéndice: Detalle de las Simulaciones tipo R y Simulaciones tipo NR de los Experimentos

Se presenta a continuación la información detallada correspondiente a las simulaciones tipo R y tipo NR definidas y analizadas en el apartado 5.2.2. Las simulaciones se refieren solamente a la ejecución con partición de 4 regiones y función de pérdida (16) del Experimento 4 de aquella sección. A diferencia de los Cuadros 8 y 9 presentados entonces, aquí se adjunta la información de los 40 diseños de dicho experimento (10 escenarios predictivos, y cuatro tamaños muestrales para cada uno de ellos), mientras entonces se eligieron algunos casos concretos ilustrativos. Consúltese en el apartado 5.2.2 el significado de las variables que aparecen en los Cuadros 10-17 a continuación. Como se explicó entonces, hemos elegido tres simulaciones representativas para cada diseño. Para cada simulación tipo NR (R), hay asociado un valor $D = |a\widehat{p} - a\widehat{q}|$. Pues bien, en concreto, se eligen las simulaciones que dieron lugar al máximo, mínimo y mediana de la serie de num valores D asociada. Denótense esas tres simulaciones por S_{MAX}, S_{MIN} y S_{MED}. Se presentan cuatro cuadros para las simulaciones tipo R y otros cuatro para las tipo NR. El primero recoge los resultados básicos de S_{MAX}, S_{MIN} y S_{MED} (num, $a\widehat{p}$ y $a\widehat{q}$). Cada uno de los otros tres ofrece la información detallada de las frecuencias relativas (observadas y bajo independencia estocástica entre datos y previsiones) para las simulaciones S_{MAX}, S_{MIN} y S_{MED}, respectivamente.

Cuadro 10. Núm. Simulac. tipo R y Pérdidas Medias

			S_{MAX}		S_{MED}		S_{MIN}	
ρ	T	num	$a\widehat{p}$	$a\widehat{q}$	$a\widehat{p}$	$a\widehat{q}$	$a\widehat{p}$	$a\widehat{q}$
0,0	10	10	0,40	1,46	0,90	1,59	0,90	1,48
0,0	25	25	0,96	1,53	1,08	1,55	1,20	1,62
0,0	50	36	1,22	1,64	1,22	1,55	1,26	1,54
0,0	100	26	1,30	1,57	1,32	1,55	1,36	1,56
0,2	10	32	0,60	1,52	0,80	1,52	0,90	1,52
0,2	25	85	0,92	1,60	1,08	1,56	1,12	1,48
0,2	50	130	1,12	1,60	1,20	1,53	1,26	1,54
0,2	100	174	1,17	1,53	1,32	1,56	1,34	1,53
0,3	10	50	0,50	1,58	0,80	1,50	0,90	1,49
0,3	25	123	0,92	1,59	1,20	1,67	1,12	1,44
0,3	50	176	1,04	1,57	1,16	1,50	1,28	1,55
0,3	100	178	1,14	1,51	1,33	1,58	1,36	1,56
0,4	10	71	0,40	1,33	1,00	1,72	0,90	1,47
0,4	25	173	0,84	1,56	1,04	1,51	1,16	1,51
0,4	50	185	0,98	1,55	1,18	1,52	1,26	1,52
0,4	100	125	1,22	1,58	1,30	1,56	1,36	1,56
0,5	10	89	0,60	1,59	0,80	1,54	0,70	1,25
0,5	25	188	0,84	1,53	1,08	1,56	1,12	1,48
0,5	50	129	1,10	1,59	1,18	1,54	1,22	1,49
0,5	100	30	1,17	1,52	1,25	1,52	1,30	1,50
0,7	10	149	0,50	1,62	0,80	1,54	0,90	1,40
0,7	25	127	0,84	1,55	0,92	1,41	0,96	1,32
0,7	50	16	0,98	1,50	1,12	1,48	1,18	1,49
0,7	100	0	-	-	-	-	-	-
0,9	10	108	0,50	1,55	0,70	1,48	0,40	0,94
0,9	25	4	0,84	1,52	1,00	1,56	1,12	1,49
0,9	50	0	-	-	-	-	-	-
0,9	100	0	-	-	-	-	-	-
0,95	10	54	0,40	1,46	0,70	1,48	0,20	0,67
0,95	25	0	-	-	-	-	-	-
0,95	50	0	-	-	-	-	-	-
0,95	100	0	-	-	-	-	-	-
-0,25	10	6	0,70	1,44	0,80	1,46	0,80	1,40
-0,25	25	2	1,08	1,49	1,16	1,56	1,16	1,56
-0,25	50	1	1,26	1,54	1,26	1,54	1,26	1,54
-0,25	100	0	-	-	-	-	-	-
-0,8	10	0	-	-	-	-	-	-
-0,8	25	0	-	-	-	-	-	-
-0,8	50	0	-	-	-	-	-	-
-0,8	100	0	-	-	-	-	-	-

Datos de simulaciones S_{MAX}, S_{MIN} y S_{MED} para Experimento 4.

Cuadro 11. Probabilidades Pérdidas Simulaciones tipo R
Simulaciones S_{MAX}

		Observadas				Bajo Indep. Datos-Prevs			
ρ	T	$\widehat{p}(0)$	$\widehat{p}(1)$	$\widehat{p}(2)$	$\widehat{p}(3)$	$\widehat{q}(0)$	$\widehat{q}(1)$	$\widehat{q}(2)$	$\widehat{q}(3)$
0,0	10	0,60	0,40	0,00	0,00	0,32	0,26	0,06	0,36
0,0	25	0,44	0,28	0,16	0,12	0,30	0,20	0,18	0,32
0,0	50	0,40	0,24	0,10	0,26	0,29	0,19	0,10	0,42
0,0	100	0,36	0,25	0,12	0,27	0,29	0,21	0,13	0,36
0,2	10	0,50	0,40	0,10	0,00	0,28	0,22	0,20	0,30
0,2	25	0,44	0,36	0,04	0,16	0,26	0,25	0,12	0,37
0,2	50	0,38	0,32	0,10	0,20	0,27	0,22	0,13	0,37
0,2	100	0,36	0,31	0,13	0,20	0,28	0,23	0,18	0,31
0,3	10	0,50	0,50	0,00	0,00	0,30	0,22	0,08	0,40
0,3	25	0,48	0,32	0,00	0,20	0,33	0,16	0,09	0,41
0,3	50	0,42	0,32	0,06	0,20	0,31	0,19	0,12	0,38
0,3	100	0,37	0,30	0,15	0,18	0,29	0,22	0,18	0,31
0,4	10	0,60	0,40	0,00	0,00	0,32	0,26	0,19	0,23
0,4	25	0,44	0,36	0,12	0,08	0,29	0,20	0,14	0,36
0,4	50	0,42	0,34	0,08	0,16	0,30	0,20	0,15	0,35
0,4	100	0,36	0,30	0,10	0,24	0,28	0,22	0,13	0,36
0,5	10	0,40	0,60	0,00	0,00	0,26	0,24	0,15	0,35
0,5	25	0,48	0,28	0,16	0,08	0,33	0,19	0,11	0,37
0,5	50	0,38	0,30	0,16	0,16	0,26	0,23	0,16	0,35
0,5	100	0,35	0,31	0,16	0,18	0,27	0,23	0,21	0,29
0,7	10	0,50	0,50	0,00	0,00	0,26	0,26	0,08	0,40
0,7	25	0,40	0,40	0,16	0,04	0,26	0,23	0,19	0,31
0,7	50	0,38	0,34	0,20	0,08	0,26	0,24	0,23	0,27
0,7	100	-	-	-	-	-	-	-	-
0,9	10	0,50	0,50	0,00	0,00	0,25	0,25	0,20	0,30
0,9	25	0,44	0,32	0,20	0,04	0,26	0,25	0,22	0,28
0,9	50	-	-	-	-	-	-	-	-
0,9	100	-	-	-	-	-	-	-	-
0,95	10	0,60	0,40	0,00	0,00	0,32	0,26	0,06	0,36
0,95	25	-	-	-	-	-	-	-	-
0,95	50	-	-	-	-	-	-	-	-
0,95	100	-	-	-	-	-	-	-	-
-0,25	10	0,50	0,40	0,00	0,10	0,28	0,26	0,20	0,26
-0,25	25	0,40	0,24	0,24	0,12	0,26	0,25	0,22	0,26
-0,25	50	0,38	0,22	0,16	0,24	0,27	0,22	0,19	0,31
-0,25	100	-	-	-	-	-	-	-	-
-0,8	10	-	-	-	-	-	-	-	-
-0,8	25	-	-	-	-	-	-	-	-
-0,8	50	-	-	-	-	-	-	-	-
-0,8	100	-	-	-	-	-	-	-	-

Estimac. basadas en frecuencias obtenidas en simulaciones S_{MAX} Expmto 4.

Cuadro 12. Probabilidades Pérdidas Simulaciones tipo R
Simulaciones S_{MED}

		Observadas				Bajo Indep. Datos-Prevs			
ρ	T	$\widehat{p}(0)$	$\widehat{p}(1)$	$\widehat{p}(2)$	$\widehat{p}(3)$	$\widehat{q}(0)$	$\widehat{q}(1)$	$\widehat{q}(2)$	$\widehat{q}(3)$
0,0	10	0,40	0,40	0,10	0,10	0,26	0,24	0,15	0,35
0,0	25	0,44	0,24	0,12	0,20	0,28	0,22	0,18	0,32
0,0	50	0,38	0,26	0,12	0,24	0,28	0,22	0,18	0,32
0,0	100	0,40	0,20	0,08	0,32	0,31	0,19	0,14	0,36
0,2	10	0,40	0,40	0,20	0,00	0,26	0,26	0,18	0,30
0,2	25	0,44	0,24	0,12	0,20	0,29	0,21	0,14	0,36
0,2	50	0,38	0,24	0,18	0,20	0,29	0,21	0,19	0,32
0,2	100	0,38	0,19	0,16	0,27	0,29	0,20	0,15	0,36
0,3	10	0,50	0,30	0,10	0,10	0,26	0,26	0,20	0,28
0,3	25	0,36	0,32	0,08	0,24	0,27	0,21	0,11	0,42
0,3	50	0,32	0,32	0,24	0,12	0,26	0,25	0,23	0,27
0,3	100	0,36	0,23	0,13	0,28	0,29	0,21	0,14	0,36
0,4	10	0,20	0,70	0,00	0,10	0,18	0,32	0,10	0,40
0,4	25	0,48	0,24	0,04	0,24	0,32	0,19	0,14	0,35
0,4	50	0,36	0,28	0,18	0,18	0,29	0,22	0,18	0,31
0,4	100	0,32	0,31	0,12	0,25	0,26	0,24	0,18	0,32
0,5	10	0,50	0,30	0,10	0,10	0,32	0,20	0,10	0,38
0,5	25	0,40	0,28	0,16	0,16	0,28	0,21	0,18	0,33
0,5	50	0,40	0,26	0,10	0,24	0,31	0,19	0,16	0,34
0,5	100	0,35	0,26	0,18	0,21	0,26	0,24	0,21	0,29
0,7	10	0,50	0,20	0,30	0,00	0,26	0,24	0,20	0,30
0,7	25	0,40	0,36	0,16	0,08	0,31	0,25	0,15	0,29
0,7	50	0,40	0,24	0,20	0,16	0,32	0,21	0,15	0,32
0,7	100	-	-	-	-	-	-	-	-
0,9	10	0,40	0,50	0,10	0,00	0,22	0,28	0,30	0,20
0,9	25	0,40	0,28	0,24	0,08	0,26	0,24	0,18	0,32
0,9	50	-	-	-	-	-	-	-	-
0,9	100	-	-	-	-	-	-	-	-
0,95	10	0,40	0,50	0,10	0,00	0,22	0,28	0,30	0,20
0,95	25	-	-	-	-	-	-	-	-
0,95	50	-	-	-	-	-	-	-	-
0,95	100	-	-	-	-	-	-	-	-
-0,25	10	0,50	0,30	0,10	0,10	0,26	0,26	0,24	0,24
-0,25	25	0,40	0,24	0,16	0,20	0,26	0,23	0,21	0,30
-0,25	50	0,38	0,22	0,16	0,24	0,27	0,22	0,19	0,31
-0,25	100	-	-	-	-	-	-	-	-
-0,8	10	-	-	-	-	-	-	-	-
-0,8	25	-	-	-	-	-	-	-	-
-0,8	50	-	-	-	-	-	-	-	-
-0,8	100	-	-	-	-	-	-	-	-

Estimac. basadas en frecuencias obtenidas en simulaciones S_{MED} Expmto 4.

Cuadro 13. Probabilidades Pérdidas Simulaciones tipo R
Simulaciones S_{MIN}

		Observadas				Bajo Indep. Datos-Prevs			
ρ	T	$\widehat{p}(0)$	$\widehat{p}(1)$	$\widehat{p}(2)$	$\widehat{p}(3)$	$\widehat{q}(0)$	$\widehat{q}(1)$	$\widehat{q}(2)$	$\widehat{q}(3)$
0,0	10	0,40	0,30	0,30	0,00	0,22	0,28	0,30	0,20
0,0	25	0,40	0,28	0,04	0,28	0,27	0,24	0,10	0,40
0,0	50	0,36	0,26	0,14	0,24	0,27	0,23	0,18	0,32
0,0	100	0,30	0,29	0,16	0,25	0,28	0,23	0,16	0,33
0,2	10	0,40	0,40	0,10	0,10	0,23	0,27	0,25	0,25
0,2	25	0,28	0,44	016	0,12	0,24	0,26	0,28	0,22
0,2	50	0,36	0,22	0,22	0,20	0,27	0,23	0,19	0,31
0,2	100	0,33	0,23	0,21	0,23	0,28	0,22	0,19	0,31
0,3	10	0,40	0,40	0,10	0,10	0,26	0,24	0,25	0,25
0,3	25	0,36	0,16	0,48	0,00	0,23	0,23	0,42	0,12
0,3	50	0,36	0,20	0,24	0,20	0,25	0,24	0,22	0,29
0,3	100	0,33	0,24	0,17	0,26	0,27	0,23	0,16	0,33
0,4	10	0,40	0,30	0,30	0,00	0,23	0,27	0,30	0,20
0,4	25	0,28	0,40	0,20	0,12	0,23	0,26	0,29	0,22
0,4	50	0,32	0,28	0,22	0,18	0,26	0,24	0,22	0,28
0,4	100	0,33	0,24	0,17	0,26	0,28	0,22	0,16	0,33
0,5	10	0,60	0,20	0,10	0,10	0,32	0,26	0,27	0,15
0,5	25	0,40	0,16	0,36	0,08	0,24	0,25	0,29	0,22
0,5	50	0,36	0,28	0,14	0,22	0,27	0,25	0,20	0,28
0,5	100	0,35	0,25	0,15	0,25	0,29	0,23	0,16	0,31
0,7	10	0,20	0,70	0,10	0,00	0,19	0,35	0,33	0,13
0,7	25	0,36	0,36	0,24	0,04	0,28	0,26	0,30	0,15
0,7	50	0,38	0,24	0,20	0,18	0,30	0,22	0,16	0,31
0,7	100	-	-	-	-	-	-	-	-
0,9	10	0,70	0,20	0,10	0,00	0,44	0,30	0,14	0,12
0,9	25	0,32	0,32	0,28	0,08	0,25	0,25	0,26	0,24
0,9	50	-	-	-	-	-	-	-	-
0,9	100	-	-	-	-	-	-	-	-
0,95	10	0,80	0,20	0,00	0,00	0,58	0,24	0,11	0,07
0,95	25	-	-	-	-	-	-	-	-
0,95	50	-	-	-	-	-	-	-	-
0,95	100	-	-	-	-	-	-	-	-
-0,25	10	0,40	0,40	0,20	0,00	0,23	0,27	0,37	0,13
-0,25	25	0,40	0,24	0,16	0,20	0,26	0,23	0,21	0,30
-0,25	50	0,38	0,22	0,16	0,24	0,27	0,22	0,19	0,31
-0,25	100	-	-	-	-	-	-	-	-
-0,8	10	-	-	-	-	-	-	-	-
-0,8	25	-	-	-	-	-	-	-	-
-0,8	50	-	-	-	-	-	-	-	-
-0,8	100	-	-	-	-	-	-	-	-

Estimac. basadas en frecuencias obtenidas en simulaciones S_{MIN} Expmto 4.

Cuadro 14. Núm. Simulac. tipo NR y Pérdidas Medias

			S_{MAX}		S_{MED}		S_{MIN}	
ρ	T	num	$a\widehat{p}$	$a\widehat{q}$	$a\widehat{p}$	$a\widehat{q}$	$a\widehat{p}$	$a\widehat{q}$
0,0	10	0	-	-	-	-	-	-
0,0	25	6	2,24	1,54	2,04	1,50	1,96	1,61
0,0	50	9	2,22	1,54	1,74	1,47	1,64	1,52
0,0	100	6	2,04	1,55	1,73	1,54	1,48	1,57
0,2	10	1	1,00	1,56	1,00	1,56	1,00	1,56
0,2	25	2	2,12	1,52	1,72	1,54	1,72	1,54
0,2	50	6	1,28	1,56	1,40	1,61	1,68	1,50
0,2	100	3	1,34	1,52	1,41	1,56	1,43	1,54
0,3	10	1	1,00	1,52	1,00	1,52	1,00	1,52
0,3	25	3	1,20	1,60	1,16	1,54	1,16	1,50
0,3	50	5	1,28	1,57	1,28	1,55	1,46	1,60
0,3	100	4	1,36	1,56	1,47	1,60	1,50	1,62
0,4	10	3	2,10	1,48	1,10	1,53	1,10	1,53
0,4	25	3	1,20	1,58	1,20	1,58	1,16	1,40
0,4	50	2	1,32	1,59	1,30	1,55	1,30	1,55
0,4	100	2	1,36	1,56	1,36	1,54	1,36	1,54
0,5	10	1	1,00	1,53	1,00	1,53	1,00	1,53
0,5	25	5	1,20	1,59	1,24	1,55	1,24	1,52
0,5	50	1	1,28	1,50	1,28	1,50	1,28	1,50
0,5	100	0	-	-	-	-	-	-
0,7	10	1	1,80	1,44	1,80	1,44	1,80	1,44
0,7	25	3	1,20	1,59	1,20	1,58	1,20	1,52
0,7	50	0	-	-	-	-	-	-
0,7	100	0	-	-	-	-	-	-
0,9	10	0	-	-	-	-	-	-
0,9	25	0	-	-	-	-	-	-
0,9	50	0	-	-	-	-	-	-
0,9	100	0	-	-	-	-	-	-
0,95	10	0	-	-	-	-	-	-
0,95	25	0	-	-	-	-	-	-
0,95	50	0	-	-	-	-	-	-
0,95	100	0	-	-	-	-	-	-
-0,25	10	8	2,50	1,58	1,00	1,36	1,80	1,66
-0,25	25	25	2,56	1,53	2,16	1,54	1,68	1,48
-0,25	50	59	2,38	1,55	2,04	1,54	1,76	1,57
-0,25	100	142	2,11	1,54	1,94	1,56	1,67	1,55
-0,8	10	56	2,90	1,58	2,70	1,73	2,20	1,59
-0,8	25	554	2,80	1,54	2,44	1,61	1,96	1,53
-0,8	50	924	2,74	1,59	2,32	1,56	1,98	1,55
-0,8	100	999	2,63	1,59	2,30	1,55	2,05	1,57

Datos de simulaciones S_{MAX}, S_{MIN} y S_{MED} para Experimento 4.

Cuadro 15. Probabilidades Pérdidas Simulaciones tipo NR
Simulaciones S_{MAX}

		Observadas				Bajo Indep. Datos-Prevs			
ρ	T	$\widehat{p}(0)$	$\widehat{p}(1)$	$\widehat{p}(2)$	$\widehat{p}(3)$	$\widehat{q}(0)$	$\widehat{q}(1)$	$\widehat{q}(2)$	$\widehat{q}(3)$
0,0	10	-	-	-	-	-	-	-	-
0,0	25	0,12	0,12	0,16	0,60	0,28	0,23	0,15	0,34
0,0	50	0,08	0,16	0,22	0,54	0,28	0,22	0,18	0,32
0,0	100	0,18	0,12	0,18	0,52	0,28	0,23	0,17	0,33
0,2	10	0,60	0,00	0,20	0,20	0,30	0,22	0,10	0,38
0,2	25	0,12	0,20	0,12	0,56	0,32	0,18	0,16	0,34
0,2	50	0,44	0,14	0,12	0,30	0,30	0,20	0,15	0,36
0,2	100	0,39	0,18	0,13	0,30	0,30	0,21	0,17	0,33
0,3	10	0,60	0,00	0,20	0,20	0,34	0,18	0,10	0,38
0,3	25	0,52	0,04	0,16	0,28	0,28	0,22	0,12	0,38
0,3	50	0,46	0,12	0,10	0,32	0,33	0,17	0,11	0,39
0,3	100	0,40	0,18	0,08	0,34	0,31	0,19	0,12	0,38
0,4	10	0,00	0,40	0,10	0,50	0,26	0,26	0,22	0,26
0,4	25	0,52	0,08	0,08	0,32	0,30	0,19	0,14	0,37
0,4	50	0,42	0,12	0,18	0,28	0,29	0,20	0,13	0,38
0,4	100	0,39	0,14	0,19	0,28	0,28	0,22	0,16	0,34
0,5	10	0,60	0,00	0,20	0,20	0,27	0,23	0,20	0,30
0,5	25	0,52	0,08	0,08	0,32	0,28	0,22	0,14	0,36
0,5	50	0,44	0,08	0,24	0,24	0,29	0,22	0,18	0,31
0,5	100	-	-	-	-	-	-	-	-
0,7	10	0,00	0,50	0,20	0,30	0,22	0,28	0,34	0,16
0,7	25	0,48	0,08	0,20	0,24	0,28	0,20	0,18	0,34
0,7	50	-	-	-	-	-	-	-	-
0,7	100	-	-	-	-	-	-	-	-
0,9	10	-	-	-	-	-	-	-	-
0,9	25	-	-	-	-	-	-	-	-
0,9	50	-	-	-	-	-	-	-	-
0,9	100	-	-	-	-	-	-	-	-
0,95	10	-	-	-	-	-	-	-	-
0,95	25	-	-	-	-	-	-	-	-
0,95	50	-	-	-	-	-	-	-	-
0,95	100	-	-	-	-	-	-	-	-
-0,25	10	0,00	0,10	0,30	0,60	0,26	0,24	0,16	0,34
-0,25	25	0,08	0,08	0,04	0,80	0,42	0,09	0,02	0,47
-0,25	50	0,10	0,10	0,12	0,68	0,34	0,17	0,10	0,40
-0,25	100	0,11	0,16	0,24	0,49	0,28	0,22	0,17	0,32
-0,8	10	0,00	0,00	0,10	0,90	0,40	0,08	0,06	0,46
-0,8	25	0,00	0,04	0,12	0,84	0,38	0,12	0,08	0,42
-0,8	50	0,04	0,02	0,10	0,84	0,33	0,17	0,08	0,42
-0,8	100	0,04	0,07	0,11	0,78	0,31	0,18	0,10	0,41

Estimac. basadas en frecuencias obtenidas en simulaciones S_{MAX} Expmto 4.

Cuadro 16. Probabilidades Pérdidas Simulaciones tipo NR
Simulaciones S_{MED}

		Observadas				Bajo Indep. Datos-Prevs			
ρ	T	$\widehat{p}(0)$	$\widehat{p}(1)$	$\widehat{p}(2)$	$\widehat{p}(3)$	$\widehat{q}(0)$	$\widehat{q}(1)$	$\widehat{q}(2)$	$\widehat{q}(3)$
0,0	10	-	-	-	-	-	-	-	-
0,0	25	0,08	0,28	0,16	0,48	0,30	0,22	0,17	0,31
0,0	50	0,20	0,26	0,14	0,40	0,32	0,21	0,15	0,32
0,0	100	0,20	0,27	0,13	0,40	0,30	0,20	0,15	0,34
0,2	10	0,60	0,00	0,20	0,20	0,30	0,22	0,10	0,38
0,2	25	0,08	0,36	0,32	0,24	0,25	0,25	0,20	0,30
0,2	50	0,42	0,08	0,18	0,32	0,27	0,21	0,17	0,36
0,2	100	0,37	0,15	0,18	0,30	0,28	0,22	0,15	0,35
0,3	10	0,60	0,00	0,20	0,20	0,34	0,18	0,10	0,38
0,3	25	0,56	0,04	0,08	0,32	0,35	0,15	0,10	0,40
0,3	50	0,42	0,12	0,22	0,24	0,28	0,22	0,17	0,33
0,3	100	0,39	0,11	0,14	0,36	0,28	0,22	0,12	0,38
0,4	10	0,50	0,10	0,20	0,20	0,30	0,20	0,17	0,33
0,4	25	0,52	0,08	0,08	0,32	0,31	0,20	0,10	0,39
0,4	50	0,42	0,12	0,20	0,26	0,29	0,21	0,17	0,33
0,4	100	0,37	0,15	0,23	0,25	0,28	0,22	0,17	0,33
0,5	10	0,60	0,00	0,20	0,20	0,27	0,23	0,20	0,30
0,5	25	0,44	0,12	0,20	0,24	0,26	0,23	0,20	0,30
0,5	50	0,44	0,08	0,24	0,24	0,29	022	0,18	0,31
0,5	100	-	-	-	-	-	-	-	-
0,7	10	0,00	0,50	0,20	0,30	0,22	0,28	0,34	0,16
0,7	25	0,52	0,08	0,08	0,32	0,27	0,24	0,12	0,37
0,7	50	-	-	-	-	-	-	-	-
0,7	100	-	-	-	-	-	-	-	-
0,9	10	-	-	-	-	-	-	-	-
0,9	25	-	-	-	-	-	-	-	-
0,9	50	-	-	-	-	-	-	-	-
0,9	100	-	-	-	-	-	-	-	-
0,95	10	-	-	-	-	-	-	-	-
0,95	25	-	-	-	-	-	-	-	-
0,95	50	-	-	-	-	-	-	-	-
0,95	100	-	-	-	-	-	-	-	-
-0,25	10	0,60	0,10	0,00	0,30	0,30	0,24	0,26	0,20
-0,25	25	0,08	0,16	0,28	0,48	0,25	0,25	0,22	0,29
-0,25	50	0,12	0,26	0,08	0,54	0,27	0,23	0,18	0,32
-0,25	100	0,20	0,14	0,18	0,48	0,29	0,21	0,16	0,35
-0,8	10	0,00	0,10	0,10	0,80	0,29	0,17	0,06	0,48
-0,8	25	0,04	0,08	0,28	0,60	0,26	0,21	0,18	0,34
-0,8	50	0,00	0,16	0,36	0,48	0,25	0,24	0,21	0,30
-0,8	100	0,09	0,12	0,19	0,60	0,30	0,20	0,15	0,35

Estimac. basadas en frecuencias obtenidas en simulaciones S_{MED} Expmto 4.

Cuadro 17. Probabilidades Pérdidas Simulaciones tipo NR
Simulaciones S_{MIN}

		Observadas				Bajo Indep. Datos-Prevs			
ρ	T	$\widehat{p}(0)$	$\widehat{p}(1)$	$\widehat{p}(2)$	$\widehat{p}(3)$	$\widehat{q}(0)$	$\widehat{q}(1)$	$\widehat{q}(2)$	$\widehat{q}(3)$
0,0	10	-	-	-	-	-	-	-	-
0,0	25	0,16	0,16	0,24	0,44	0,30	0,15	0,20	0,35
0,0	50	0,16	0,32	0,24	0,28	0,27	0,24	0,18	0,31
0,0	100	0,39	0,10	0,15	0,36	0,29	0,21	0,14	0,36
0,2	10	0,60	0,00	0,20	0,20	0,30	0,22	0,10	0,38
0,2	25	0,08	0,36	0,32	0,24	0,25	0,25	0,20	0,30
0,2	50	0,14	0,30	0,30	0,26	0,25	0,25	0,25	0,25
0,2	100	0,39	0,10	0,20	0,31	0,30	0,21	0,16	0,34
0,3	10	0,60	0,00	0,20	0,20	0,34	0,18	0,10	0,38
0,3	25	0,48	0,08	0,24	0,20	0,27	0,24	0,20	0,29
0,3	50	0,40	0,10	0,14	0,36	0,26	0,24	0,13	0,37
0,3	100	0,37	0,12	0,15	0,36	0,28	0,21	0,10	0,40
0,4	10	0,50	0,10	0,20	0,20	0,30	0,20	0,17	0,33
0,4	25	0,48	0,08	0,24	0,20	0,35	0,21	0,14	0,30
0,4	50	0,42	0,12	0,20	0,26	0,29	0,21	0,17	0,33
0,4	100	0,37	0,15	0,23	0,25	0,28	0,22	0,17	0,33
0,5	10	0,60	0,00	0,20	0,20	0,27	0,23	0,20	0,30
0,5	25	0,44	0,04	0,36	0,16	0,26	0,24	0,22	0,28
0,5	50	0,44	0,08	0,24	0,24	0,29	0,22	0,18	0,31
0,5	100	-	-	-	-	-	-	-	-
0,7	10	0,00	0,50	0,20	0,30	0,22	0,28	0,34	0,16
0,7	25	0,48	0,08	0,20	0,24	0,28	0,22	0,19	0,31
0,7	50	-	-	-	-	-	-	-	-
0,7	100	-	-	-	-	-	-	-	-
0,9	10	-	-	-	-	-	-	-	-
0,9	25	-	-	-	-	-	-	-	-
0,9	50	-	-	-	-	-	-	-	-
0,9	100	-	-	-	-	-	-	-	-
0,95	10	-	-	-	-	-	-	-	-
0,95	25	-	-	-	-	-	-	-	-
0,95	50	-	-	-	-	-	-	-	-
0,95	100	-	-	-	-	-	-	-	-
-0,25	10	0,10	0,40	0,10	0,40	0,25	0,21	0,17	0,37
-0,25	25	0,12	0,24	0,48	0,16	0,24	0,26	0,28	0,22
-0,25	50	0,12	0,36	0,16	0,36	0,25	0,24	0,20	0,31
-0,25	100	0,20	0,32	0,09	0,39	0,29	0,22	0,15	0,35
-0,8	10	0,00	0,30	0,20	0,50	0,23	0,27	0,18	0,32
-0,8	25	0,12	0,28	0,12	0,48	0,29	0,22	0,16	0,33
-0,8	50	0,12	0,24	0,18	0,46	0,27	0,23	0,18	0,32
-0,8	100	0,16	0,16	0,15	0,53	0,28	0,22	0,16	0,34

Estimac. basadas en frecuencias obtenidas en simulaciones S_{MIN} Expmto 4.

Referencias

[1] Amstrong, J.S.y Fildes, R. (1995). Correspondence on the Selection of Error Measures for Comparison among Forecasting Methods, *Journal of Forecasting* 14, 67-71.

[2] Ash, J.C.K., Smyth, D.J y Heravi, S.M. (1998). Are OECD Forecasts Rational and Useful?: a Directional Analysis, *International Journal of Forecasting* 14, 381-391.

[3] Baillie, R.T., Diebold, F.X. y Engle, R.F., et al. (1993). On the Limitations of Comparing Mean Square Forecast Errors, *Journal of Forecasting* 12, 639-667.

[4] Barlow, R.E. et al. (1972). Statistical Inference under Order Restrictions: the Theory and Application of Isotonic Regression. Wiley Series in Probability and Mathematical Statistics (New York).

[5] Birchenall, C.R., Jessen, H. y Osborn, D.R. (1996). Predicting US Business Cycle Regimes, *Journal of Business and Economic Statistics*, 17, 313-323.

[6] Clements, M.P. y Hendry, D.F. (1993). On the Limitations of Comparing Mean square Forecast Errors, *Journal of Forecasting* 12, 617-637.

[7] Cumby, R.E. y Modest, D.M. (1987). Testing for Market Timing Ability, *Journal of Financial Economics* 19, 169-189.

[8] Greer, M. (2003). Directional Accuracy Tests of Long-Term Interest Rate Forecasts, *International Journal of Forecasting* 19, 291-298.

[9] Henriksson, R.D., Merton y R.C. (1981). On Market Timing and Investment Perfomance. II: statistical procedures for evaluating forecasting skills, *Journal of Business* 54, 513-533.

[10] Joutz, F. y Stekler, H.O. (2000). An Evaluation of the Predictions of the Federal Reserve, *International Journal of Forecasting* 16, 17-38.

[11] Kolb, R.A. y Stekler, H.O. (1996). How well do Analysts Forecast Interest Rates?, *Journal of Forecasting* 15, 385-394.

[12] Leitch, G. y Tanner, J.E. (1995). Professional Economic Forecasts: Are they Worth their Costs?, *Journal of Forecasting* 14, 143-157.

[13] Merton, R.C. (1981). On Market Timing and Investment Performance. I: an Equilibrium Theory of Value for Market Forecasts, *Journal of Businsess* 54, 363-406.

[14] Mills, T.C. y Pepper, G.T. (1999). Assessing the Forecasters: an Analysis of the Forecasting Records of the Treasury, the London Business School and the National Insititute, *International Journal of Forecasting* 15, 247-257.

[15] Novales, A. (1993). Econometría. McGraw-Hill (Madrid).

[16] Oller, L. y Bharat, B. (2000). The Accuracy of European Growth and Inflation Forecasts, *International Journal of Forecasting* 16, 293-315.

[17] Pesaran, M.H. y Timmermann, A. (1992). A Simple Nonparametric Test of Predictive Performance, *Journal of Business and Economic Statistics* 10, 461-465.

[18] Pesaran, M.H. y Timmermann, A. (1994). A Generalisation of the Non-parametric Henriksson-Merton Test of Market Timing, *Economics Letters* 44, 1-7.

[19] Pons, J. (2000). The Accuracy of IMF and OECD Forecasts for G7 Countries, *Journal of Forecasting* 19, 53-63.

[20] Robertson, T. y Wright, F.T. (1981). Likelihood Ratio Tests for and against a Stochastic Ordering between Multinomial Populations, *The Annals of Statistics*, vol. 9, n^o 6, 1248-1257.

[21] Schnader, M.H. y Stekler, H.O. (1990). Evaluating Predictions of Change, *The Journal of Business* 63, 1, 99-107.

[22] Stekler, H.O. (1994). Are Economic Forecasts Valuable?, *Journal of Forecasting* 13, 495-505.

CAPÍTULO 2

CONTRASTES ESTADÍSTICOS PARA COMPARAR CAPACIDAD PREDICTIVA ENTRE DOS CONJUNTOS DE PREVISIONES, BAJO FUNCIÓN DE PÉRDIDA DISCRETA

1. Introducción

La comparación entre la capacidad predictiva de dos modelos o dos conjuntos de previsiones es un asunto clave en la práctica económica y sobre el que se ha generado extensa literatura en los últimos años. Desde que Diebold y Mariano (1995) presentaron su test (DM, en adelante), éste se ha convertido en la referencia fundamental respecto a contrastes estadísticos sobre igualdad de bondad predictiva entre dos conjuntos alternativos de previsiones. Pese a su carácter asintótico, el test DM posee buenas propiedades estadísticas también en muestras finitas, es aplicable con una amplia gama de funciones de pérdida y es robusto al incumplimiento de supuestos estándar sobre los errores de previsión (media cero, normalidad, no autocorrelación, no correlación contemporánea entre las dos series de errores). Diebold y Mariano (1995) muestran que su contraste es bastante preciso en tamaño, salvo cuando la longitud de la muestra es muy pequeña (inferior a 20 datos). Estos resultados fueron obtenidos implementando el test DM con una función de pérdida igual al cuadrado del error de previsión.

El trabajo de Diebold y Mariano (1995) se concibió bajo el supuesto de que la varianza de las previsiones generadas solo procedía del carácter aleatorio de la variable a predecir y no de la estimación, es decir, bajo el supuesto de que, si se utilizan modelos de previsión, los parámetros que intervienen en ellos son conocidos. Recientemente, muchos otros trabajos han adaptado el contraste DM para casos en que las previsiones descansan en estimaciones de modelos y la incertidumbre paramétrica debe ser tenida en cuenta. Sin embargo, este artículo se enmarca en la línea original de Diebold y Mariano, y su extensión incorporando incertidumbre paramétrica constituirá el contenido de nuestro siguiente capítulo.

Nuestro trabajo expone dos aportaciones de interés. Por un lado, sugerimos un tipo de función de pérdida peculiar (ya introducida en el Capítulo 1) para aplicar los contrastes de comparación de capacidad predictiva, bajo la que puede implementarse el propio DM. Mostramos sus ventajas en términos de robustez del test a la presencia de atípicos en la muestra de errores de previsión. Además, constatamos que las propiedades del contraste DM en muestras finitas en aquellas circunstancias predictivas analizadas por Diebold y Mariano (1995) para función de pérdida igual al error cuadrático se conservan si se usa la función que proponemos. La segunda aportación del capítulo es la verdaderamente relevante. La naturaleza no continua de dicha función nos permite derivar dos contrastes (válidos exclusivamente cuando la función de pérdida es ésta), que pueden constituir una alternativa a DM en este contexto, y exploramos sus propiedades de tamaño y potencia en muestras cortas (entre 8 y 48 datos), bajo distintos supuestos. Uno de los tests (que denotaremos por Mult2), constituye la aportación fundamental del capítulo. Como se verá, Mult2 posee un tamaño más exacto que DM en casi cualquiera de las situaciones simuladas, y mayor potencia en ciertos casos. Por el contrario, no es robusto a la presencia de autocorrelación en las pérdidas y tiene un coste computacional relativamente elevado.

Como es bien sabido, el test DM contrasta la hipótesis $E(d_t) = 0$, siendo $d_t = g(y_t, v_{t,2}) - g(y_t, v_{t,1})$, y_t el dato de la variable a prever para el periodo t y $v_{t,1}$, $v_{t,2}$ las previsiones alternativas para dicho dato y g una función de pérdida. Aunque DM no exige que g satisfaga ninguna condición particular, la práctica casi generalizada es aplicar el test bajo una función cuadrática en el error de previsión, es decir, $g(y_t, v_{t,i}) = (y_t - v_{t,i})^2 = e_{t,i}^2$.[1] Es decir, la información del par $(y_t, v_{t,i})$ se condensa en $e_{t,i}$ y la penalización viene dada por el cuadrado de éste. Lo primero puede implicar una pérdida de información crucial, mientras lo segundo podría conducir al deterioro de las propiedades del test de igualdad de capacidad predictiva si aparecen valores extremos en la muestra, tal y como veremos más adelante.

El error de previsión pierde los signos de y_t y $v_{t,i}$ y, por lo tanto, su empleo como medida básica de la bondad de predicción será muy inapropiado en todos aquellos casos en que la comparación entre ambos signos fuera relevante. En muchas aplicaciones económicas, la variable a prever (y_t) es la tasa de

[1] A lo largo del capítulo será habitual que nos refiramos a dicha función simplemente como "función cuadrática".

variación de una magnitud e interesa evaluar fundamentalmente si la previsión acierta el signo de dicha tasa o incluso del cambio en la tasa (dirección del cambio), mientras la proximidad o lejanía entre previsión y dato se valora solo de forma somera. La rentabilidad de muchos activos financieros y las magnitudes de Contabilidad Nacional (PIB, por ejemplo) son variables que encajan perfectamente con el enfoque expuesto.

Pues bien, la función de pérdida que nosotros sugerimos pretende responder a este tipo de situaciones (otras buenas propiedades conceptuales de esta función de pérdida pueden consultarse en la Introducción de la Tesis o en el apartado 3.2.2 del Capítulo 1). Nos referiremos a ella a partir de ahora por "función de pérdidas discreta". Aunque dicha función se definirá con rigor en la sección 2, trataremos de adelantar ahora los conceptos con un ejemplo sencillo. Imaginemos que y_t es el cambio en la tasa de variación del PIB de una economía y que el analista desea valorar sus previsiones en cuanto a capacidad de predicción correcta del signo de y_t y, en segundo plano, en cuanto a magnitud del error. A favor de esta forma de valorar las previsiones se pronuncian explícitamente, por ejemplo, Bierchenall et al (1996) y Leicht y Tanner (1995). Por su parte, Ash, Smyth y Heravi (1998), Schnader y Stekler (1990) y Stekler (1994), entre otros, realizan análisis sobre predicción macroeconómica también bajo este enfoque.

Una posible función de pérdidas discreta g en dicho ejemplo vendría dada por la siguiente matriz:

		$v_{t,i}$			
		G+	P+	P-	G-
y_t	G+	0	1	2	3
	P+	1	0	2	3
	P-	3	2	0	1
	G-	3	2	1	0

(G1)

, donde G y P simbolizan cambios en la tasa de variación del PIB de tamaño "grande" y "pequeño", respectivamente (el usuario deberá fijar un valor crítico que separe ambos niveles), y +, -, los signos de dicho cambio. Cualquier otra especificación coherente con la lógica del caso concreto sería igualmente válida. Por ejemplo, también sería razonable pensar que la magnitud del error solo es relevante si el signo ha sido correctamente previsto (en cuyo caso los valores 2 y 3 de posiciones adyacentes en la matriz deberían ser iguales entre sí) o podría interesar introducir asimetrías (prever cambios positivos en la tasa de crecimiento del PIB cuando fueron negativos puede ser más grave que lo contrario, o viceversa). La sensibilidad de las propiedades estadísticas de los tests DM y Mult2 a la especificación concreta de la función g será el tema de la sección 5. En nuestro contexto, definir g significa especificar la partición del dominio de los datos y previsiones (en el ejemplo, G+, G-, P+ y P-) y asignar las penalizaciones para cada cuadrante.

Como se adelantó arriba, el empleo de la habitual función $g(e_{t,i}) = e_{t,i}^2$ puede no ser adecuado incluso en casos en que sintetizar en $e_{t,i}$ la información correspondiente al par $(y_t, v_{t,i})$ sí sea apropiado. En primer lugar, porque el signo del error puede ser trascendente y los errores negativos ser más penalizables que los positivos. Y en segundo lugar, porque, bajo dicha función, la aparición de valores atípicos en una de las series de errores de previsión distorsiona completamente el resultado del test utilizado. Sin embargo, alternativamente, se podrían usar funciones de pérdida discretas, del tipo a (G1), y se evitarían ambos problemas. Para ello, basta especificar la partición sobre el dominio de errores de previsión en vez de sobre datos y previsiones, asignando valores numéricos de pérdida a las regiones generadas, en vez de a los cuadrantes. Es decir, en lugar de la matriz mostrada antes se tendrían clasificaciones, por ejemplo, del tipo:

$e_{t,i}$			
G+	P+	P-	G-
2	0	1	3

valorando asimétricamente los errores positivos y negativos, o bien

$e_{t,i}$		
G+	P	G-
1	0	2

(G2)

donde P simboliza errores de cuantía "pequeña", y la asimetría según signos se introduce solo si el error es "grande".

En la sección 2 se comparan resultados de tamaño y potencia para el test DM bajo dos implementaciones: por un lado, con la función cuadrática habitual $g = e_{t,i}^2$ y, alternativamente, utilizando el sistema de clasificación discreta que denotábamos por (G2). Las Tablas 1 y 2 se corresponden con experimentos en los que $e_{t,i} \sim N(0,1)$ pero donde se introduce sistemáticamente un valor atípico igual a 10 en la muestra $e_{t,1}$. La probabilidad de rechazo de la hipótesis nula del test DM disminuye drásticamente por la aparición del valor atípico si se emplea la función cuadrática, mientras, como era esperable, no se altera en absoluto si se usó la discreta. Por otro lado, las Tablas 3 y 4 exponen los cambios que se producen en el tamaño y potencia de DM cuando la distribución generadora de los errores de previsión fue una $t - Student$ con 3 grados de libertad, de colas más gruesas que las de la $N(0,1)$. En la práctica, es más probable que los errores se ajusten a una distribución de este tipo que a una Normal. Ahora, el tamaño de DM converge a niveles inferiores al teórico y la potencia empeora notablemente respecto al caso gaussiano si se usa la función cuadrática, mientras las alteraciones de estas propiedades son mínimas cuando la función empleada es (G2). En su artículo, Diebold y Mariano defendían que el test DM era robusto en muestras finitas a desviaciones respecto a normalidad y presentaban pruebas en dicho sentido utilizando una distribución $t(6)$ para generar las series $e_{t,i}$. Sin embargo, con distribuciones cuya probabilidad de generar valores extremos es algo mayor que en la $t(6)$, el resultado de Diebold y Mariano no se verifica. Fijémonos que en los dos ejemplos planteados, el comportamiento de DM se deteriora no solo en muestras cortas sino también en longitudes muy elevadas. Dell'Aquila y Ronchetti (2004) mostraron evidencia en esta misma línea.

Todo lo anterior hace razonable el empleo de funciones de clasificación g discretas, para la comparación de capacidad predictiva entre dos conjuntos de previsiones y éstas, a su vez, abren puertas a la construcción de nuevos contrastes alternativos a DM, aprovechando el carácter de variables aleatorias discretas que van a tomar ahora las pérdidas. Probamos con dos tests (RV-p y Mult2) y evaluamos sus propiedades estadísticas en muestras cortas, comparativamente con el test DM y con dos contrastes no paramétricos estándar válidos en muestras finitas, los tests de Signos y de Wilcoxon, todos ellos implementados con función de pérdida g discreta. Llevamos a cabo un estudio de Monte Carlo muy similar al realizado por Diebold y Mariano (1995) al presentar su test DM y los resultados son favorables al contraste Mult2. Si las series de errores de previsión $e_{t,i}$ no son autocorreladas −el caso habitual si el horizonte de previsión es uno−, el tamaño de Mult2 es el más exacto de todos los tests y su potencia, la mayor (Tablas 5 y 7 de la sección 4). Si existe autocorrelación, la superioridad de Mult2 respecto a DM es indiscutible mientras la longitud muestral sea inferior a 20 datos, y desaparece para longitudes mayores (Tablas 6 y 8 de la sección 4). En general, si el usuario de tests de comparación de habilidad predictiva se halla en una aplicación en la que una función de pérdida discreta (con una partición no fina) es adecuada y el número de datos de la muestra es pequeño, digamos no superior a 50, procede el empleo de Mult2 en vez de DM. Esto es especialmente aconsejable si las pérdidas están autocorreladas y la muestra de previsiones no contiene más de 20 datos. No obstante, también en todos estos casos existen razones favorables al uso de DM, relacionadas con la implementación de los tests. Por un lado, el cálculo de Mult2 es más complejo computacionalmente (aunque presentaremos una versión asintótica del test, con buenos resultados) y, por otro, en caso de que exista autocorrelación de orden r en las series de pérdidas, Mult2 debería implementarse usando contrastes múltiples bajo la cota de Bonferroni, como sugirieron en otro contexto Campbell y Ghysels (1995), lo que implica ejecutar el test $r+1$ veces (una para cada una de las submuestras no autocorreladas). Afortunadamente, en situaciones de autocorrelación, el uso de dicho procedimiento alivia enormemente el coste computacional del contraste.

El artículo se estructura como sigue: En la sección 2 se definen las funciones discretas que usaremos y se aporta evidencia a su favor respecto a la función cuadrática habitual. Los tests estadísticos a evaluar son presentados en la sección 3 y los resultados de dicha evaluación, en términos de tamaño y potencia, se muestran en la sección 4. La sensibilidad de estos resultados a la especificación concreta de la función de pérdida es objeto de análisis en la sección 5. Además, dicha sección adjunta un breve estudio sobre el sesgo en tamaño que se produce en los tests implementados con función de pérdidas discreta bajo ciertos escenarios de predicción peculiares si la muestra es muy corta. Finalmente, en la sección 6 se exponen las principales conclusiones del trabajo.

2. Aplicación de la función de pérdida discreta en contextos de comparación de capacidad predictiva

2.1. Motivación de las pérdidas discretas. Robustez ante atípicos y ante no normalidad en los errores de previsión

Como se comentó en la Introducción, el cuadrado del error de previsión constituye una pobre medida de bondad predictiva en la mayoría de las aplicaciones económicas. Con frecuencia, uno desea penalizar las previsiones según su capacidad para adelantar correctamente los signos de las variables objeto de predicción, en cuyo caso obviamente no debe sintetizarse dato y previsión por la diferencia entre ambos, o bien se pueden querer valorar asimétricamente errores de la misma magnitud pero de signo contrario. Hay muchos ejemplos en la literatura donde la métrica para evaluar previsiones se desmarca del error cuadrático medio. Schnader y Stekler (1990), Stekler (1994), Leitch y Tanner (1991, 1995), Chinn y Meese (1991), Ash, Smyth y Heravi (1998) y Greer (2003), por ejemplo, se enfocan hacia la dirección del cambio en la variable, West, Edison y Cho (1993) enfatizan en criterios basados en utilidad y Clements y Hendry (1993) incluso proponen una nueva medida, el error cuadrático medio generalizado (GMSE).

Por todo esto, pensamos que los tests para contrastar igual precisión entre dos conjuntos de previsiones que compiten entre sí deben establecerse sobre funciones de pérdida cuyos inputs sean el dato y la previsión, ambos, y solo como caso particular, éstos pueden reducirse a un solo input a través del error de previsión. Éste es el enfoque bajo el que se construye el contraste DM, y, sin embargo, su uso extendido es precisamente con función cuadrática $g(y_t, v_{t,i}) = e_{t,i}^2$. Probablemente, la razón es que no es inmediato diseñar funciones continuas que se adapten correctamente al tipo de valoración que el usuario tiene en mente. Por el contrario, mucho más sencillo en la práctica parece definir funciones discretas del tipo a las que hemos expuesto en la Introducción. Esencialmente, se trata de (a) especificar una partición de m regiones para el dominio de datos y previsiones y (b) asignar a cada uno de los m^2 cuadrantes un valor numérico que determina la penalización. En la próxima sección se definen formalmente (a) y (b).

Por su flexibilidad, este mecanismo permite especificar adecuadamente el criterio conceptual de valoración de las previsiones para cada situación concreta y, si m es pequeño, hacerlo con rapidez. Tal y como se sugería en la Introducción, en muchas aplicaciones prácticas de evaluación de previsiones económicas este método parece muy razonable, especialmente allá donde prever correctamente el signo de la variable sea esencial y, a la vez, la magnitud del fallo también deba ser tenida en cuenta, pero sin exigirse gran precisión. Por ejemplo, en muchos contextos podría ser perfectamente admisible valorar igual dos previsiones de cambio en la tasa de crecimiento anual del PIB del +3 % y +4 %, respectivamente, cuando el incremento verdadero fue de +0,5 %. En esta línea, Leitch y Tanner (1995) utilizan el test DM con una función de pérdida de este tipo para comparar previsiones respecto del cambio en la tasa de crecimiento del GNP de EEUU, realizadas por dos modelos alternativos. Si nos salimos del marco de la macroeconomía, existen innumerables aplicaciones en las que las funciones de pérdida continua ni siquiera tienen sentido, mientras las discretas son totalmente adecuadas: siempre que la variable sobre la que se hace previsión no sea de carácter cuantitativo. Si se trata de prever colores, tallas, marcas, tipos de individuos u objetos, etc, la asignación de pérdidas solo puede ser realizarse de forma natural a través de una función discreta. Ejemplos posibles son la previsión metereológica sobre si el día va a ser "soleado", "nuboso", "lluvioso", etc, o el diagnóstico de tipos de enfermedades. Más aún, incluso sería posible asignar pérdidas no numéricas, y, sin embargo, nuestro enfoque seguirá funcionando igualmente.

Las cualidades conceptuales de la función de pérdida discreta resaltadas hasta ahora habían sido mencionadas en el Capítulo 1 de la Tesis. Pero, además de ellas, este tipo de funciones presenta otro punto favorable adicional que no poseen muchas funciones continuas: robustez a la aparición de valores atípicos en alguna de las muestras. A continuación se adjuntan los resultados de varios experimentos de simulación en los que se evalúan las propiedades estadísticas del test DM utilizando dos funciones de pérdida alternativas, la función cuadrática habitual $g(y_t, v_{t,i}) = e_{t,i}^2$ y una función discreta, donde los errores de previsión se clasifican en G+ ($e > +1$), G- ($e < -1$) y P ($e \in [-1, +1]$) y se penalizan según (G2). El diseño de los experimentos es el siguiente:

– Experimento 1 (Tabla 1): Los errores de previsión se han generado por $e_{t,i} \sim N(0, 1)$, no presentan autocorrelación ($\theta = 0$) ni correlación contemporánea ($\rho = 0$). Para cada tamaño muestral, se ejecuta el test DM, utilizando un nivel de significación $\alpha = 0{,}05$ y, tras 10000 repeticiones, se estima el tamaño del

test. Posteriormente, el ejercicio se repite añadiendo un atípico de valor 10 en la muestra 1 (es decir, se hace la sustitución $e_{1,1} = 10$ antes de aplicar las funciones de pérdidas). Se trata de comprobar el efecto que la aparición de un valor extremo en una de las muestras ejerce sobre la probabilidad de rechazo de la hipótesis nula por parte del test DM, según la función de pérdida usada.

– Experimento 2 (Tabla 2): Se vuelve a ejecutar el experimento anterior, pero generando los errores por $e_{t,i} \sim N(0, \sigma_i^2)$, con $\sigma_1^2 = 1{,}5 > 1 = \sigma_2^2$. Ahora, por tanto, la hipótesis nula de que la capacidad predictiva asociada a los dos conjuntos de errores de previsión es la misma es falsa, y, por tanto, se está evaluando potencia. Se trata de comprobar el efecto que la aparición de un atípico tiene sobre dicha propiedad en el test DM, según la función de pérdida empleada.

– Experimento 3 (Tablas 3 y 4): Ahora, se vuelven a realizar experimentos como 1 y 2 pero, en vez de introducir un valor atípico de forma determinista, el experimento se lleva a cabo con otra distribución generadora, la $t-$Student, con mayor probabilidad de producir valores extremos que la $N(0, 1)$. En concreto, se usa $e_{t,i} \sim t(3)$. Las series de errores procedentes de $t(3)$ (cuya varianza teórica es 3) se estandarizan para asegurar varianza 1, como en Diebold y Mariano (1995). En el ejercicio de potencia, se emplea de nuevo $\sigma_1^2 = 1{,}5$. Se usa nivel de significación $\alpha = 0{,}05$. La prueba consiste en chequear las alteraciones en el tamaño y potencia de DM cuando los errores no proceden de una distribución normal.

En los ejercicios de potencia, estimaremos dos medidas alternativas de ésta. Por un lado, la medida habitual, es decir, la probabilidad de rechazar la hipótesis nula cuando ésta es falsa. Por otro lado, estimaremos la "potencia ajustada a tamaño" (utilizaremos la notación SAP, por "size-adjusted power"), criterio extendido en la literatura estadística actual. Como es bien sabido, dicha medida evita la distorsión que presenta la comparación entre contrastes en términos de potencia cuando alguno de ellos es sesgado en tamaño (un test que rechaza su hipótesis nula pocas veces cuando es cierta (sesgo por defecto en tamaño), tiende a hacer lo mismo cuando es falsa (infraestimando potencia), y viceversa, resultando una imagen irreal favorable en potencia a los que rechazan más de lo debido (sesgados en tamaño por exceso)).

Cuando se quiere medir la potencia de un mismo test en dos contextos de evaluación diferentes, lo correcto es utilizar la medida habitual de potencia, ya que refleja la probabilidad de rechazo que el usuario del test se encontraría en la práctica. Sin embargo, si lo que se pretende es comparar la potencia de dos contrastes, la comparación debe realizarse en términos de SAP, para evitar las distorsiones que se acaban de mencionar.

Los resultados de estas tres pruebas ilustran claramente que la robustez del test DM a la aparición de valores extremos en alguna de las muestras de pérdidas depende del tipo de función g empleada. Si ésta es $g = e_{t,i}^2$, las decisiones del test se distorsionan notablemente al introducir valores atípicos. En cambio, éstos no producen ningún efecto si g es de tipo discreto.

La Tabla 1 muestra cómo, en el caso de que la muestra de errores de previsión contenga un valor atípico igual a 10, la probabilidad de rechazo del test DM se aproxima a cero si se empleó g cuadrática, incluso en muestras muy largas. Y cuando la hipótesis nula es falsa, el test no lo detecta prácticamente nunca, a menos que la longitud de la muestra (que denotaremos por T) sea muy elevada: según los resultados de la Tabla 2, así ocurre mientras $T < 256$. En principio, se podría haber esperado que, al aparecer un valor atípico, la potencia del test se modificara al alza, es decir, que el valor atípico ayudara a identificar las series de errores como diferentes. Sin embargo, ocurre lo contrario, porque la varianza del estadístico $\overline{d}$ utilizado en el contraste aumenta extraordinariamente.[2] Es decir, el contraste DM no es capaz de detectar la falsedad de la hipótesis nula en caso de que algún error de previsión de una de las dos series tome un valor extremo, si se usó la habitual función $g = e_{t,i}^2$.[3]

En la misma línea se sitúan los resultados del Experimento 3, expuestos en las Tablas 3 y 4. Es razonable pensar que una distribución de colas gruesas se adapte mejor que una distribución Normal a las series de errores de previsión que se pueden obtener en la mayor parte de las aplicaciones de predicción, ya que

[2] En realidad, lo que ocurre es que la calidad de la estimación de la varianza del estadístico $\overline{d}$ se deteriora notablemente. Consideremos que el estadístico de contraste es $\overline{d}$ y su distribución es $N(0, V(\overline{d}))$, en vez de la implementación equivalente habitual (estadístico de contraste $\overline{d}V(\overline{d})^{-1/2}$ y distribución $N(0, 1)$). La estimación de $V(\overline{d})$ tiende a ser mucho mayor que la verdadera e implícitamente, por tanto, se está utilizando una distribución con varianza muy superior a la correcta, de modo que el valor crítico se sitúa erróneamente en una posición mucho más alejada del centro de la distribución que la correcta, sesgando a la baja el tamaño y la potencia del test.

[3] Por supuesto, al corregir la potencia por el tamaño empírico en el test DM con función de pérdida cuadrática, se recuperan prácticamente los niveles de potencia del caso sin valores atípicos, porque el nivel crítico utilizado en la implementación SAP del contraste se adapta a los resultados de los ejercicios de tamaño (véase Apéndice 1). Sin embargo, esto no es relevante. La medición adecuada en esta situación es la probabilidad de rechazo, por la razón explicada anteriormente en el texto.

la segunda asigna probabilidad prácticamente nula a valores muy extremos (véase Cuadro 1). Utilizando una distribución $t(3)$ para generar los errores, el tamaño de DM converge a un nivel aproximadamente dos puntos inferior al teórico si el test se implementa con g cuadrática, y la potencia se reduce bruscamente respecto al caso $N(0,1)$. Con función cuadrática, la potencia empeora incluso usando la medición SAP, y es inferior también a la obtenida con función discreta. Por el contrario, ninguna propiedad se ve afectada significativamente si g fue de tipo discreto (solamente se observa una moderada pérdida de potencia).

Diebold y Mariano (1995) habían probado que el tamaño del test DM con $g = e_{t,i}^2$ no sufría cambios cuando los errores procedían de una $t(6)$ en vez de una Normal, lo que se interpretó como robustez a la distribución generadora de los mismos. Sin embargo, como mostramos en la Tabla 3, este resultado no es general y basta una distribución que produzca valores extremos con probabilidad un poco más alta que $t(6)$ (véase Cuadro 1) para que el tamaño de DM implementado con función cuadrática ya no converja al teórico y, además, la potencia disminuya considerablemente (sobre esta propiedad no se pronunciaron Diebold y Mariano en su artículo). Experimentos y conclusiones muy similares fueron obtenidas por Dell'Aquila y Ronchetti (2004), solo que ellos presentan el problema como inherente al test y no a la función de pérdida empleada y, del mismo modo, abogan por soluciones sobre la definición del contraste (por otro lado, difíciles de llevar a la práctica), mientras nosotros mostramos cómo el uso de funciones discretas permite salvar las propiedades del test de forma sencilla.

No obstante, si bien la introducción de funciones de pérdida discretas implica robustez a cambios en la distribución de los errores de previsión, también tiene una contrapartida. Como puede observarse en las Tablas 2 y 4, cuando la distribución de $e_{t,i}$ es $N(0,1)$ (es decir, si la probabilidad de valores extremos es tan pequeña como la asociada a una $N(0,1)$), la potencia del test DM usando g discreta es bastante inferior a la obtenida con g cuadrática. Pero lo contrario ocurre si la probabilidad de atípicos se asemeja más bien a la asociada a una $t(3)$, conclusión que consideramos de notable interés.

Tabla 1: Probabilidad Rechazo Test DM
$e_{t,i} \sim N(0,1)$, $\theta = \rho = 0$
$\alpha = 5\,\%$. 10000 repeticiones

	T	g discreta	g cuadrática
Sin atípico	8	12,1	9,5
	16	7,4	6,5
	32	5,8	5,8
	64	5,2	5,6
	128	5,2	5,2
	256	4,9	4,6
	512	4,8	4,9
Atípico $e_{1,1} = 10$	8	12,5	0,0
	16	7,5	0,0
	32	6,6	0,0
	64	5,7	0,0
	128	5,3	0,0
	256	5,2	0,1
	512	5,1	0,6

Tabla 2: Potencia Empírica DM
$e_{t,i} \sim N(0, \sigma_i^2)$, $\sigma_1^2 = 1,5$, $\sigma_2^2 = 1$, $\theta = \rho = 0$
$\alpha = 5\,\%$. 10000 repeticiones

		Potencia Sin Ajustar a Tamaño		Potencia Ajustada a Tamaño (SAP)	
Sin atípico	T	g discreta	g cuadrática	g discreta	g cuadrática
	8	14,0	17,6	10,1	11,6
	16	15,8	21,6	13,0	18,9
	32	20,2	31,0	18,6	28,9
	64	28,4	49,3	28,2	48,3
	128	44,1	73,9	44,0	74,2
	256	68,0	94,3	69,2	94,2
	512	91,1	99,8	91,1	99,8
Atípico $e_{1,1} = 10$	8	17,6	0,0	9,5	8,2
	16	18,4	0,0	11,7	13,1
	32	21,8	0,0	16,7	18,9
	64	29,8	0,2	27,8	36,6
	128	46,0	11,1	43,0	62,0
	256	68,6	72,2	68,4	90,3
	512	91,6	99,4	90,6	99,7

Tabla 3: Tamaño Empírico DM
$E(e_{t,i}) = 0$, $\sigma_i^2 = 1$, $\theta = \rho = 0$
$\alpha = 5\,\%$. 10000 repeticiones

	T	g discreta	g cuadrática
$N(0,1)$	8	12,4	9,4
	16	7,8	7,1
	32	6,4	6,0
	64	5,8	5,7
	128	5,5	5,1
	256	5,0	4,4
	512	5,3	4,9
$t(3)$	8	9,5	6,7
	16	6,7	4,6
	32	6,5	4,0
	64	5,7	3,3
	128	5,1	2,8
	256	5,2	3,4
	512	5,1	3,4

Tabla 4: Potencia Empírica DM
$E(e_{t,i}) = 0$, $\sigma_1^2 = 1,5$, $\sigma_2^2 = 1$, $\theta = \rho = 0$
$\alpha = 5\,\%$. 10000 repeticiones

		Potencia Sin Ajustar a Tamaño		Potencia Ajustada a Tamaño (SAP)	
$N(0,1)$	T	g discreta	g cuadrática	g discreta	g cuadrática
	8	14,7	17,6	8,4	10,7
	16	16,4	21,6	12,7	16,4
	32	19,9	30,8	17,8	26,7
	64	28,2	48,3	27,8	45,4
	128	43,5	73,0	42,3	72,4
	256	67,3	94,2	67,5	94,6
	512	90,2	99,9	90,0	99,9
$t(3)$	8	11,1	12,8	10,6	9,6
	16	14,5	12,7	10,8	11,7
	32	17,6	15,1	15,3	16,2
	64	24,6	20,0	23,0	21,9
	128	37,1	26,7	37,1	29,3
	256	57,1	36,9	56,6	38,8
	512	82,5	48,7	82,5	50,9

Cuadro 1. Probabilidades[4] valores extremos

	$P(e < x)$			
	$x = -1$	$x = -3$	$x = -5$	$x = -10$
$N(0,1)$	0,159	0,0013	$2{,}8\times10^{-5}$	$\simeq 0$
$t(6)$	0,133	0,0051	0,0004	$\simeq 0$
$t(3)$	0,091	0,0069	0,0016	0,0001

2.2. Definición formal de la función de pérdidas discreta

Aunque los ejemplos y comentarios hasta este punto del artículo ya dan una idea del tipo de función g en la que estamos pensando, procederemos a continuación a una definición formal general. La definición de este tipo de función ya fue introducida en el Capítulo 1 de la Tesis, pero la recordaremos ahora y, además, la extenderemos para que sirva para evaluar no solo el par dato-previsión, sino también directamente el error de previsión.

Sea y_t el dato en el periodo t, $v_{t,i}$ la previsión para t correspondiente al conjunto i-ésimo y $e_{t,i} = y_t - v_{t,i}$, cuyos dominios de definición son D_y, D_v y D_e, e imponemos la restricción $D_y = D_v$.[5] Sea $z_{t,i} = g(y_t, v_{t,i})$ la pérdida en t asociada a la previsión del conjunto i-ésimo, g será una función definida del siguiente modo: $g : D_y \times D_v \to A = \{a_1, a_2, ..., a_J\}$, donde a_i puede o no ser numérica y J será generalmente pequeño. Es decir, g debe verificar solamente que el número de pérdidas posibles a asignar a un par $(y_t, v_{t,i})$ sea finito. No obstante, en principio, g se corresponderá con una de las dos siguientes especificaciones:

[1] Basada en la partición del dominio de datos/previsiones (para casos donde no se desea condensar la información del par $(y_t, v_{t,i})$ en $e_{t,i}$):

a) Se realiza una partición del dominio $D_y = D_v$ en m regiones $r_1, r_2, ..., r_m$ y el conjunto de regiones se denota por R_y, es decir, $R_y = \{r_1, r_2, ..., r_{m_y}\}$. De este modo, el dominio bidimensional de los datos y previsiones $D_y \times D_v$ ha quedado particionado en m^2 "cuadrantes", y cada par $(y_t, v_{t,i})$ tendrá asociado uno de ellos.

[4] Las probabilidades que aquí se adjuntan para distribuciones t se corresponden con distribuciones $t(6)$ y $t(3)$ estandarizadas posteriormente, para mantener varianza 1.

[5] En la práctica, no parece razonable que el dominio de las previsiones no sea igual al de los datos. Dado esto, la condición se impone esencialmente para simplificar notación.

b) Se define una función $\varphi : D_y \rightarrow R_y$ que asigna una región a cada dato/previsión.

c) Finalmente, la función de pérdida g queda definida por $g : R_y \times R_v \rightarrow A$ y asigna una pérdida a cada uno de los *cuadrante* en que haya quedado particionado implícitamente el espacio $D_y \times D_v$ (se verifica, por tanto, la restricción $J \leq m_y^2$).[6] La matriz (G1) de la Introducción es un ejemplo de este tipo de construcción de la función g. Las pérdidas se denotarán por $z_{t,i}$ y quedan definidas de forma obvia a través de las funciones φ y g: $z_{t,i} = g(\varphi(y_t), \varphi(v_{t,i}))$.

De este modo, siempre que los elementos de A sean números reales, se está definiendo implícitamente la variable aleatoria discreta correspondiente a las pérdidas $Z_i = g(\varphi(Y), \varphi(V_i))$, cuyo soporte (finito) es A, siendo Y y V_i las variables aleatorias correspondientes a los datos y las previsiones del conjunto i-ésimo, cuyas realizaciones son y_t, $v_{t,i}$. Las realizaciones de Z_i son, obviamente, las pérdidas $z_{t,i}$.

[2] Basada en la partición del dominio de errores de previsión (para casos donde se desea condensar la información del par $(y_t, v_{t,i})$ en $e_{t,i}$):

a) Ahora, la partición se establece sobre D_e y el conjunto de regiones resultantes de la partición es $R_e = \{r_1, ..., r_{m_e}\}$.

b) La función φ es ahora $\varphi : D_e \rightarrow R_e$, asignando una región a cada error de previsión.

c) La función g queda ahora definida como $g : R_e \rightarrow A$ y asigna una pérdida a cada *región* en que ha quedado particionado explícitamente el espacio D_e (forzosamente, se verificará $J \leq m_e$). La especificación (G2) de la Introducción es un ejemplo de este tipo de construcción de la función g. Las pérdidas se denotarán por $z_{t,i}$ y quedan definidas de forma obvia a través de las funciones φ y g: $z_{t,i} = g(\varphi(e_{t,i}))$.

De este modo, siempre que los elementos de A sean números reales, se está definiendo implícitamente la variable aleatoria discreta correspondiente a las pérdidas $Z_i = g(\varphi(E_i))$, cuyo soporte (finito) es A, siendo $E_i = Y - V_i$ variable aleatoria cuyas realizaciones son $e_{t,i}$. Las realizaciones de Z_i son, obviamente, las pérdidas $z_{t,i}$.

Nótese que los pasos a)-c) de [1] y [2] constituyen la formalización de los puntos (a) y (b) de la definición intuitiva de función de pérdida discreta que se adelantó al principio del apartado 2.1.

2.3. Definición de la función de comparación entre pérdidas discretas

En el apartado precedente se acaba de definir la función discreta que sugerimos para valorar las previsiones $v_{t,i}$ de cada conjunto de previsiones i-ésimo, *por separado*. Sin embargo, el tema que realmente se trata en este trabajo es la comparación entre dos conjuntos de previsiones alternativos y esto implica que, además de preclasificar los pares $(y_t, v_{t,i})$ en una primera fase asignando una pérdida $z_{t,i}$, el paso posterior debe ser comparar esas pérdidas $z_{t,1}$, $z_{t,2}$ *entre sí*, siguiendo algún mecanismo de comparación, normalmente una función $f(z_{t,1}, z_{t,2})$. Dicho mecanismo constituye el objeto de este apartado.

Obviamente, cualquier test sobre igualdad de capacidad predictiva entre dos conjuntos de previsiones define un modo de hacer la comparación entre $z_{t,1}$ y $z_{t,2}$. Por ejemplo, el test DM identifica $f(z_{t,1}, z_{t,2})$ directamente con la diferencia $d_t = z_{t,2} - z_{t,1}$, el test de Signos usa simplemente el signo de d_t, mientras el de Wilcoxon incluye en su valoración también el rango de $|d_t|$. Pues bien, todos los contrastes que presentaremos en la sección 3 como alternativa a los habituales en la literatura utilizarán implícitamente la misma especificación de $f(z_{t,1}, z_{t,2})$ *que volverá a ser una función no continua, aplicable sobre pérdidas* $z_{t,1}$, $z_{t,2}$ *que hayan sido generadas por una función* g *discreta*. Dicha especificación es extremadamente simple y constituye una extensión natural de la definición de g del apartado 2.2:

$f : A \times A \rightarrow B = \{b_1, b_2, ..., b_K\}$, donde b_i sí debe ser numérica y, forzosamente, se verificará $K \leq J^2$. Dado que el número de pérdidas posibles J es limitado, f puede especificarse como una matriz C de dimensiones $J \times J$ tal que $C(i, j)$ sea la valoración para el caso $(z_{t,1} = a_i,\ z_{t,2} = a_j)$, y vendrá dada por un elemento de B. Esta valoración debe reflejar si la pérdida $z_{t,1}$ "es preferida" a $z_{t,2}$ (lo denotamos por $z_{t,1} \succ z_{t,2}$) o viceversa, y en qué medida. Si las pérdidas eran numéricas, se verificará que $z_{t,1} \succ z_{t,2} \Leftrightarrow z_{t,1} < z_{t,2}$. Los siguientes puntos son fundamentales para nuestra construcción de f:

1) Por convención, las valoraciones para casos $z_{t,1} \succ z_{t,2}$ tendrán siempre signo positivo y viceversa.

2) C deberá verificar las siguientes propiedades: (a) $C(i, i) = 0$, $i = 1, ..., J$, y (b) $C(i, j) = -C(j, i)$, $i \neq j$. Estas propiedades son coherentes con la lógica de la mayor parte de mecanismos de comparación de previsiones en que podamos pensar.

[6] En caso de ser pérdidas numéricas, los valores en A se ordenan crecientemente, $a_1 < a_2 < ... < a_J$. Lo razonable es que $a_1 = 0$ y, por tanto, el resto de pérdidas sean estrictamente positivas.

3) De lo anterior, se deduce inmediatamente que

(a) el conjunto B puede escribirse como $\{-b_q, -b_{q-1}, ..., -b_1, 0, +b_1, ..., +b_{q-1}, +b_q\}$, siendo $q = \frac{K-1}{2}$;

(b) K siempre será impar y verificará $K \leq J^2 - (J-1) = J(J-1) + 1$.

4) Notaremos con zz_t la valoración o pérdida final asociada al par $(z_{t,1}, z_{t,2})$. Es decir, sean $z_{t,1} = a_i$, $z_{t,2} = a_j$, entonces $zz_t = f(z_{t,1}, z_{t,2}) = C(i,j)$.

5) Quizá el ejemplo más natural de función f cuando las pérdidas $z_{t,i}$ son numéricas es $f(z_{t,1}, z_{t,2}) = z_{t,2} - z_{t,1}$, que es exactamente la misma función f utilizada por el test DM. Sin embargo, éste es un caso particular y nuestra especificación es más versátil.

Implícitamente, se ha definido una variable aleatoria discreta $ZZ = f(g(\varphi(Y), \varphi(V_1)), g(\varphi(Y), \varphi(V_2)))$, cuyo soporte B es finito, y cuyas realizaciones se denotan por zz_t. La finitud del soporte de ZZ será un elemento clave para el diseño de nuestros contrastes, en la sección 3.2. En el caso más frecuente de que los elementos en A sean números reales, la definición de ZZ es simplemente $ZZ = f(Z_1, Z_2)$.

Aunque el modo de proceder descrito en los apartados precedentes 2.2 y 2.3 es sencillo, vamos a ilustrarlo con un caso concreto. Por ejemplo, dados dos conjuntos de previsiones y los datos asociados, el procedimiento podría ser:

1^o) Especificación de función g de pérdidas:

Se clasifican los datos y las previsiones de *cada* conjunto en base a tres regiones y, por lo tanto, resultan 9 cuadrantes, asignando pérdidas del conjunto $A = \{0, 1, 2\}$ a cada par $(y_t, v_{t,i})$, lo que genera las pérdidas $z_{t,1}$ y $z_{t,2}$. La clasificación vendría dada por:

		$v_{t,i}$: G+	P	G-
y_t	G+	0	1	2
	P	1	0	1
	G-	2	1	0

En este caso, la especificación ha sido basada en el dominio de datos/previsiones y no en el de los errores. Obviamente, $m_y = 3$ y $J = 3$.

Es interesante resaltar que, aunque A se ha definido por valores numéricos, simplemente representa un orden de preferencias. Podría haberse definido por $A = (A_1, A_2, A_3)$, donde $A_1 \succ A_2 \succ A_3$.

2^o) Especificación de la función f de comparación de pérdidas:

Ahora se toman los pares $(z_{t,1}, z_{t,2})$ y se asignan pérdidas a cada uno. Resultan $J^2 = 9$ cuadrantes,[7] a los que hay que asociar valores numéricos, constituyendo así la matriz C. Supóngase que se adopta el mismo método de comparación entre pérdidas que en DM, es decir, $zz_t = z_{t,2} - z_{t,1}$. En tal caso, resultan $K = 5$ valores posibles en la función de comparación, a saber: $B = \{-2, -1, 0, +1, +2\}$:

		$z_{t,2}$: 0	1	2
$z_{t,1}$	0	0	+1	+2
	1	-1	0	+1
	2	-2	-1	0

(F2)

Por ejemplo, si una previsión concreta del conjunto 1 se clasificó en el paso 1^o) como "2" (la mayor pérdida posible) y la previsión del mismo periodo pero del conjunto 2 se clasificó como "0" (menor pérdida posible), la asignación a ese par es -2, la pérdida del cuadrante (3,1).

[7] El hecho de que el número de cuadrantes asociados a las funciones g y f haya resultado el mismo es solo casual. Ocurre porque se asignó una pérdida a_i distinta a cada región de la partición (de modo que $J = m_y$), lo que no tiene porqué suceder en general.

3. Contrastes sobre igualdad de capacidad predictiva

Dadas dos muestras de pérdidas $\{g(y_1, v_{1,1}), ..., g(y_T, v_{T,1})\}$ y $\{g(y_1, v_{1,2}), ..., g(y_T, v_{T,2})\}$ asociadas a los dos conjuntos de previsiones, en teoría, el contraste respecto a la existencia de diferencias en su capacidad predictiva podría plantearse, fundamentalmente, a través de tests no paramétricos de homogeneidad de muestras o de tests que comparan la posición en la recta real de las distribuciones de probabilidad de las variables aleatorias $g(Y, V_1)$ y $g(Y, V_2)$ que generaron dichas muestras. Entre los tests de homogeneidad, destacan los tests Chi-Cuadrado, Kolmogorov-Smirnov y Kruskal-Wallis. El problema de dichos contrastes es que requieren que las muestras a comparar sean independientes entre sí, condición que rara vez va a cumplirse en la mayoría de las aplicaciones que nos interesan, puesto que las fuentes de información utilizadas para la construcción de la previsión son probablemente comunes a ambos predictores, al menos parcialmente. El mismo asunto afecta a los tests de comparación de la posición de dos poblaciones que admiten diferentes longitudes en las muestras, por ejemplo, el test de la Mediana y el de Mann-Whitney. Sin embargo, si los datos son apareados, tal y como es el caso que nos ocupa, los conocidos tests no paramétricos de Signos y Wilcoxon, válidos en muestras finitas, pueden aplicarse sobre la variable $d = g(Y, V_2) - g(Y, V_1)$, contrastando la posición en la recta real de la distribución F_d, posición que caracterizan a través de la mediana. Este mismo enfoque es el adoptado por Diebold y Mariano (1995) para comparar capacidad predictiva de dos conjuntos de previsiones, solo que su test (DM) es válido solo asintóticamente, caracteriza la posición de F_d a través de la esperanza matemática en vez de la mediana y, en teoría, es robusto a la presencia de autocorrelación. Al evitar el requisito de independencia entre las muestras, seleccionamos estos tres tests habituales en la literatura, Signos, Wilcoxon y DM, como buenos candidatos para contrastar igualdad de capacidad predictiva de dos conjuntos de previsiones. Teóricamente, pueden aplicarse con cualquier función de pérdida, así que los usaremos con la función de pérdida discreta g presentada en 2.2. Hasta ahora, no existen valoraciones sobre su funcionamiento con dicha función. Nos interesa especialmente comprobar si las buenas propiedades en muestras finitas que Diebold y Mariano (1995) mostraron para su test asintótico DM usando una pérdida cuadrática se conservan al usar la función discreta.

Al aplicar sobre datos y previsiones una función de pérdidas discreta, $g(Y, V_1)$ y $g(Y, V_2)$ son las variables aleatorias discretas Z_1 y Z_2 que introdujimos en 2.2. Esta característica permite construir contrastes específicos basados en la distribución Multinomial para este contexto, que compararemos con los tres mencionados, que son generales en cuanto a la función de pérdida admisible. Los dos tests que propondremos contrastarán hipótesis sobre la distribución F_{ZZ} de la variable $ZZ = f(Z_1, Z_2)$, que no tiene por qué ser de la forma $ZZ = Z_2 - Z_1$ (aunque en la práctica será lo habitual), a diferencia de lo que ocurre en Signos, Wilcoxon y DM. Igual que éstos, nuestros contrastes eluden el problema de correlación entre muestras, al condensar las dos variables de pérdidas en una sola. El primero de los tests propuestos sigue el enfoque de DM, en el sentido de contrastar la posición de F_{ZZ} en la recta real, identificada por la esperanza matemática. Expondremos una versión en muestras finitas (Mult2), tratando de construir numéricamente la distribución F_{ZZ}, de la que no existe expresión analítica, aunque no llegamos a obtener la distribución exacta. Además, presentaremos la versión asintótica de dicho test (Mult2-aprx), que, como es lógico, difiere muy poco de DM cuando se usa la especificación $ZZ = Z_2 - Z_1$ (si las pérdidas no presentan autocorrelación, estos dos tests solo se diferencian en la estimación de la varianza del estadístico de contraste). El segundo contraste que proponemos es paramétrico, y contrasta ciertas restricciones sobre el vector paramétrico que caracteriza la distribución de un vector aleatorio de frecuencias que está directamente relacionado con ZZ. Nuestra propuesta es simplemente aplicar en este contexto la versión asintótica estándar del test de Razón de Verosimilitudes.

3.1. Contrastes estándar de la literatura

Presentamos a continuación el test DM, y las versiones de Signos y Wilcoxon para contrastar igualdad de capacidad predictiva entre dos conjuntos de previsiones. Se exponen para una función de pérdida genérica, pero en nuestros experimentos de Monte Carlo serán implementados siempre bajo funciones discretas.

3.1.1. Test de Diebold y Mariano (1995)

Diebold y Mariano (1995) presentaron un test de carácter asintótico, válido en condiciones muy generales, para contrastar igualdad en precisión entre dos conjuntos de previsiones. El test puede aplicarse aun incumpliéndose los supuestos habituales respecto a los errores de previsión: media cero, ausencia de autocorrelación y de correlación contemporánea, y distribución Normal. Más aún, admite una amplia gama de funciones de pérdida para evaluar las predicciones (podrían ser, en particular, no cuadráticas, no simétricas, e incluso no continuas), que, además, utilicen como inputs el dato y la previsión, y no necesariamente el error de previsión. El procedimiento estadístico DM es el siguiente:

Sean las previsiones "competidoras" $\{v_{t,1}\}_{t=1}^{T}$, $\{v_{t,2}\}_{t=1}^{T}$, sean $\{y_t\}_{t=1}^{T}$ los datos asociados a dichas previsiones, y sea la función de pérdida $z_{t,i} = g(y_t, v_{t,i})$. Se construye el diferencial de pérdida $f(z_{t,1}, z_{t,2}) = d_t = z_{t,2} - z_{t,1}$,[8] obteniéndose la secuencia temporal $\{d_t\}_{t=1}^{T}$. Usando resultados estándar, Diebold y Mariano proponen contrastar $H_0^{(DM)} = E(d_t) = 0$ a través de:

$$S_1 = \frac{\overline{d}}{\sqrt{\frac{2\pi \widehat{f}_d(0)}{T}}} \overset{a}{\sim} N(0,1), \tag{1}$$

donde $\overline{d}$ es la media muestral del diferencial de pérdidas y $f_d(0) = \frac{1}{2\pi} \sum_{\tau=-\infty}^{\infty} \gamma_d(\tau)$ la densidad espectral del diferencial de pérdidas en la frecuencia cero, siendo, a su vez, $\gamma_d(\tau)$ la autocovarianza de orden τ del diferencial.

El contraste será unilateral o bilateral, según se tengan a prioris sobre cuál es el conjunto de previsiones superior. Dada la definición elegida para d_t, si el rechazo de $H_0^{(DM)}$ se produce por la cola derecha (izquierda) indica superioridad de las previsiones del conjunto uno (dos).

La estimación de $f_d(0)$ constituye el aspecto más problemático en la implementación del test. Una estimación consistente se obtiene de la suma ponderada de las autocovarianzas muestrales: $2\pi \widehat{f}_d(0) = \sum_{\tau=-(T-1)}^{T-1} l\left(\frac{\tau}{L(T)}\right) \widehat{\gamma}_d(\tau)$, siendo $l\left(\frac{\tau}{L(T)}\right)$ el "lag-window" y $L(T)$ el retardo de truncamiento y con $\widehat{\gamma}_d(\tau) = \frac{1}{T} \sum_{t=|\tau|+1}^{T} (d_t - \overline{d})(d_{t-|\tau|} - \overline{d})$. La elección del retardo de truncamiento y el "lag-window" es un asunto complejo (véase Diebold y Mariano (1995) para una exposición detallada sobre el tema), pero la práctica extendida consiste en utilizar $l\left(\frac{\tau}{L(T)}\right) = 1$ si $\left|\frac{\tau}{L(T)}\right| \leq 1$, y $l\left(\frac{\tau}{L(T)}\right) = 0$, en otro caso ("lag-window" rectangular), mientras $L(T)$ se suele definir por $L(T) = h - 1$, siendo h el horizonte de previsión.[9] Ésta es la implementación que eligen Diebold y Mariano (1995) en sus experimentos de simulación y será también la que utilizaremos en los nuestros.

Como se dijo antes, el test DM tiene la ventaja de su gran versatilidad, mientras sus limitaciones son esencialmente dos:

a) La distribución de S_1 es conocida solo asintóticamente. Sus propiedades en muestras finitas deben ser puestas a evaluación. Diebold y Mariano (1995) presentaron evidencia sobre el tamaño del test bajo $g(y_t, v_{t,i}) = e_{t,i}^2$, resultando éste considerablemente por encima del teórico en muestras de longitud inferior a 32 previsiones, pero aproximadamente correcto si la muestra es mayor. Ahora, queremos evaluar el

[8] Se ha definido la diferencia en este orden (pérdida 2 menos pérdida 1) por coherencia con el orden utilizado en la especificación de la función f en el apartado 2.3.

[9] La elección de $L(T)$ se basa en el argumento de que, de acuerdo a la descomposición de Wold para un proceso estocástico regular y lineal, los errores de previsión óptimos siguen una estructura MA($h-1$), y, por tanto, su autocorrelación será de orden $h-1$, como máximo.

tamaño y la potencia del test DM si g es de tipo discreto y concentrándonos en muestras pequeñas, entre 8 y 48 datos. Comprobaremos si los resultados de DM en este contexto pueden ser mejorados por los contrastes que se proponen en el apartado 3.2.

b) La estimación de $f_d(0)$ empleando el "lag-window" rectangular mencionado arriba no garantiza no negatividad. Esto significa que el contraste puede no ser aplicable en algunos casos, a menos que se recurra a definiciones alternativas del "lag-window" que, por su grado de complejidad, harían el test mucho menos atractivo en la práctica.

Modificación de Harvey, Leybourne y Newbold (HLN). Harvey, Leybourne y Newbold (1997) proponen una ligera modificación del test DM, válida solo para corregir aquéllas implementaciones del test que estimaran $f_d(0)$ usando "lag-window" rectangular y $L(T) = h-1$. La nueva versión del contraste DM, que denotaremos por HLN, logra algunas mejoras en el tamaño del test en muestras pequeñas, y consiste en sustituir el estadístico S_1 de (1) por el siguiente estadístico:

$$S_1^* = \left(\frac{T+1-2h+T^{-1}h(h-1)}{T}\right)^{1/2} S_1 \tag{2}$$

3.1.2. Test de Signos y Test de Wilcoxon

Diebold y Mariano (1995) evaluaron su contraste DM frente a dos tests no paramétricos estándar, Signos y Wilcoxon, que, tal y como hemos comentado en la introducción a esta sección, pueden ser perfectamente utilizados para contrastar igualdad de precisión entre dos muestras de previsiones, aplicándolos sobre la secuencia $\{d_t\}_{t=1}^{T}$ que se definió para el test DM. La hipótesis nula sería $H_0^{(SW)} = med(d_t) = 0$ (med: mediana), difiriendo ligeramente de $H_0^{(DM)}$. La implementación de estos tests es muy simple:

Sea $\{d_t\}_{t=1}^{T'}$ una secuencia de diferenciales igual a la anterior pero habiendo eliminado los elementos $d_t = 0$. Sea ahora una variable indicatriz $I(d_t) = 1$ si $d_t > 0$, mientras $I(d_t) = 0$ en caso contrario, entonces los estadísticos de Signos (S_2) y Wilcoxon (S_3) quedan especificados por:

$$\begin{aligned} S_2 &= \sum_{t=1}^{T'} I(d_t) \sim B(T', 0{,}5) \\ S_3 &= \sum_{t=1}^{T'} I(d_t) rank(|d_t|), \end{aligned} \tag{3}$$

donde $B(n,p)$ simboliza distribución Binomial de parámetros n y p (en este caso, $p = P(I(d_t) = 1)$) y $rank(|d_t|)$ denota la posición del dato en una reordenación creciente de la muestra de valores absolutos de los diferenciales. El estadístico S_3 no sigue una distribución de probabilidad estándar pero sus valores críticos en muestras finitas han sido tabulados. Véase que la especificación de la función f de comparación entre pérdidas viene dada por $f(z_{t,1}, z_{t,2}) = I(d_t)$ en el test de Signos y por $f(z_{t,1}, z_{t,2}) = I(d_t)rank(|d_t|)$ en el de Wilcoxon, siendo $d_t = z_{t,2} - z_{t,1}$.

La versión asintótica de los tests es:[10]

$$\begin{aligned} S_{2a} &= \frac{S_2 - 0{,}5T'}{\sqrt{0{,}25T'}} \overset{a}{\sim} N(0,1), \\ S_{3a} &= \frac{S_2 - \frac{T(T'+1)}{4}}{\sqrt{\frac{T'(T'+1)(2T'+1)}{24}}} \overset{a}{\sim} N(0,1). \end{aligned} \tag{4}$$

Los contrastes serán unilaterales o bilaterales, según existan aprioris sobre cuál es el conjunto de previsiones superior, y el rechazo de la hipótesis nula por la derecha (izquierda) tendrá la misma interpretación que se explicó en el caso del test DM.

[10] La demostración de $E(S_3) = \frac{T(T+1)}{4}$ y $V(S_3) = \frac{T(T+1)(2T+1)}{24}$ es sencilla y puede verse, por ejemplo, en Novales, "Estadística y Econometría", capítulo 12.

Los tests de Signos y Wilcoxon tienen la ventaja de ser válidos en muestras finitas y son aplicables, igual que DM, bajo correlación contemporánea, errores no gaussianos y de media no nula. Sin embargo, presentan algunos aspectos desfavorables:

1) El test de Signos solo discrimina entre los conjuntos de previsiones por el número de veces que la pérdida de uno fue mayor que la del otro, pero no por la magnitud de su diferencia. Será, por tanto, poco potente (esto es un defecto general de un test de Signos, en cualquier contexto).

2) Las distribuciones de probabilidad exacta y asintótica de ambos tests son correctas si la secuencia $\{d_t\}_{t=1}^{T'}$ es *iid*. Es decir, los tests de Signos y Wilcoxon requieren aleatoriedad, lo que, en este contexto sobre predicción, se puede traducir esencialmente por no autocorrelación en el diferencial de pérdidas d_t. La solución en caso de autocorrelación es la realización de contrastes múltiples sobre submuestras no autocorreladas usando la cota de Bonferroni para definir el tamaño de los tests parciales, método utilizado por Diebold y Mariano (1995) para la aplicación de Signos y Wilcoxon en sus experimentos de simulación. En el cuarto apartado de 3.2.5 se detalla el procedimiento, que es muy sencillo.

3) La derivación de la distribución para muestras finitas de S_2 y S_3 requiere el cumplimiento del supuesto de continuidad de la variable d_t. En primer lugar, porque debe verificarse que $P(I(d_t) > 0) = 1/2$ bajo la hipótesis nula, igualdad que puede no cumplirse si d_t es una variable discreta y $P(d_t = 0) \neq 0$. Pues bien, precisamente en nuestro contexto, d_t no es continua y el suceso $\{d_t = 0\}$ es muy frecuente, especialmente si las previsiones de los dos conjuntos están correladas. Para solucionar el problema, la muestra original $\{d_t\}_{t=1}^{T}$ debe "limpiarse" de elementos nulos, y aplicarse luego el contraste sobre el resto de la muestra, cuya longitud pasa a ser T', tal y como se mencionó antes. Éste es el procedimiento aconsejado en todos los manuales de estadística y el que emplearemos nosotros en nuestras simulaciones. De este modo, las distribuciones derivadas siguen siendo correctas para la muestra de longitud T'. Sin embargo, debe advertirse que, aun siendo ésta la implementación más recomendable de los tests, el efecto de la misma puede ser grave en nuestro contexto de variable d_t discreta, al generar una reducción de longitud muestral muy significativa, disminuyendo notablemente la fiabilidad del resultado de estos contrastes. Por ejemplo, en los ejercicios de tamaño que presentaremos en la próxima sección, el valor de $P(d_t = 0)$ es 0,51, 0,62 y 0,83, para los tres niveles de correlación entre errores de previsión que se considerarán en dichos experimentos: $\rho = 0, 0{,}5, 0{,}9$, respectivamente, siendo ρ el coeficiente de correlación entre $e_{t,1}$ y $e_{t,2}$. Esto significa que, incluso en casos de previsiones con un grado de correlación intermedio, muestras de diferenciales cuyas longitudes originales eran 8, 16 y 24, quedan reducidas a apenas 3, 6 y 9 datos, respectivamente. Esta limitación no será visible en los resultados de simulación, debido a la aplicación del método de "aleatorización", que se explicará en el segundo apartado de 3.2.5, pero estará presente en la práctica.

4) Además, la continuidad de d_t era requerida en la derivación de la distribución de muestras finitas del estadístico S_3 en un segundo aspecto. Garantizaba la imposibilidad de que existieran dos observaciones muestrales con el mismo valor, condición fundamental para la correcta derivación de la distribución de S_3, ya que las probabilidades tabuladas para S_3 se calculan suponiendo que cada observación de la muestra recibe un rango $rank(d_t)$ igual a uno de los números naturales entre 1 y T'. Cuando la condición no se cumple y se producen "empates", el consejo de los manuales de estadística es asignar un rango promedio a las observaciones empatadas. Éste será el procedimiento que emplearemos nosotros para el cálculo de S_3, pero puede suponer que el contraste no tenga un tamaño exacto, ya que la distribución teórica de contraste no es exactamente la verdadera distribución de S_3. Fijémonos que en nuestro contexto, en primer lugar, la probabilidad $P(d_t = 0)$ tendrá un valor relativamente elevado, pero, además, entre los T' diferenciales restantes, la probabilidad $P(d_t = d_s)$ tampoco será, en general, igual a cero, siendo $t \neq s$ y $d_t \neq 0$. Es decir, habrá diferenciales no nulos iguales entre sí, y, por tanto, aparecerán situaciones de "empates" con relativa frecuencia. Por ello, pese a ser, en principio, válido para muestras finitas, el test de Wilcoxon puede presentar cierto sesgo en tamaño en nuestro contexto. Comprobaremos el efecto de este asunto en nuestros experimentos de simulación.

5) Por último, además de continuidad, el test de Wilcoxon requiere simetría en la distribución de d_t para que la derivación del contraste sea correcta. Aunque en nuestros ejercicios de simulación, no se incumplirá dicha condición, supone una restricción adicional para su aplicación práctica.

3.2. Contrastes propuestos

Como ya se anticipó en la introducción a esta sección, el uso de la función de pérdida discreta hace que las variables aleatorias $g(Y, V_1)$ y $g(Y, V_2)$ generadoras de las pérdidas tengan un soporte finito, en cuyo caso las denotábamos por Z_1 y Z_2.[11] Así, la variable $ZZ = f(Z_1, Z_2)$ que las compara también será una variable aleatoria discreta con soporte finito y, como veremos a continuación, las frecuencias asociadas a la muestra $(zz_1, ..., zz_T)$ tendrán una distribución conocida. Todas estas circunstancias facilitan el diseño de tests de comparación de capacidad predictiva específicos en el contexto de las funciones de pérdida discretas. Primero, presentamos un contraste sobre la posición de la distribución F_{ZZ} en la recta real, cuya hipótesis será $E(ZZ) = 0$, interpretándose ésta por "igualdad de capacidad predictiva entre ambos conjuntos de previsiones", y lo presentamos en sus versiones para muestras finitas y asintótica. No derivamos la distribución exacta, pero la aproximación propuesta se revelará como superior a la proporcionada por la distribución asintótica. Segundo, se propone un contraste paramétrico sobre el vector de parámetros caracterizador de la distribución de las frecuencias asociadas a las pérdidas, en forma de test de Razón de Verosimilitudes, especificado en su versión asintótica.

Antes de pasar a la exposición de los tests, introducimos el contexto y definiciones necesarias para que dicha exposición sea fluida.

3.2.1. Notación, planteamiento e hipótesis

Distribución de las frecuencias asociadas a las comparaciones de pérdidas La especificación de la función f para comparar pérdidas que anunciamos en el apartado 2.3 permite un planteamiento estadístico sencillo para nuestros tests. Dados los conjuntos de datos $\{y_t\}_{t=1}^T$ y las previsiones competidoras $\{v_{t,1}\}_{t=1}^T$ y $\{v_{t,2}\}_{t=1}^T$, se utiliza un sistema de clasificación con función de pérdida g discreta (apartado 2.2), que genera dos secuencias de pérdidas $\{z_{t,1}\}_{t=1}^T$ y $\{z_{t,2}\}_{t=1}^T$. A su vez, éstas son comparadas entre sí según una función f del tipo expuesto en el apartado 2.3, que asigna a cada par $(z_{t,1}, z_{t,2})$ uno de los K valores numéricos de un conjunto B (donde, por convención, los valores asociados a casos $z_{t,1} \succ z_{t,2}$ serán positivos, y viceversa), generándose la secuencia final $\{zz_t\}_{t=1}^T$. Pues bien, sea $b = (-b_q, -b_{q-1}, ..., -b_1, 0, +b_1, ..., +b_{q-1}, +b_q)$ un vector fila conteniendo esos K valores de B ordenados de forma creciente ($q = \frac{K-1}{2}$) y denotando por p_i la probabilidad $P(ZZ = b(i))$, entonces la secuencia $\{zz_t\}_{t=1}^T$ puede considerarse como las T realizaciones de un experimento aleatorio con K sucesos posibles excluyentes entre sí, cuyas probabilidades son $p_1, p_2, ..., p_K$. Por tanto, el vector de frecuencias $(n_1, n_2, ..., n_K)'$ sigue una *distribución multinomial* con parámetros T y $p = (p_1, p_2, ..., p_K)$.

Verdaderamente, la afirmación anterior solo es correcta si las realizaciones del experimento, y, por tanto, las variables zz_t, son mutuamente independientes, lo que podría no ocurrir. Por eso, en tales casos, la implementación de nuestros tests seguirá el procedimiento explicado en el cuarto apartado de 3.2.5.

En adelante, utilizaremos el subíndice s para denotar el elemento central de los vectores b y p. Es decir, $b_s = 0$, y, por tanto, $p_s = P(ZZ = 0)$. Obviamente, $s = \frac{K+1}{2} = q + 1$.

Hipótesis En el contexto que acabamos de presentar, existen esencialmente dos posibles formalizaciones de la hipótesis nula de "igualdad de capacidad predictiva en ambos conjuntos":

a) La primera posible hipótesis es "la esperanza matemática de la distribución de ZZ es cero":

$$H_0^{(1)} = E(ZZ) = bp = 0.$$

En este caso, la igualdad predictiva se interpreta como "la pérdida esperada de los dos conjuntos de previsiones es la misma". Obviamente, si $ZZ = d = Z_2 - Z_1$, se trata de la misma hipótesis del test DM. La hipótesis alternativa, que denotaremos por $H_1^{(1)}$, puede definirse como bilateral ($H_1^{(1d)} = E(ZZ) = bp \neq 0$) o unilateral ($H_1^{(1u)} = E(ZZ) = bp > 0$ ó $H_1^{(1u)} = E(ZZ) = bp < 0$), según los a prioris sobre cuál de los conjuntos de previsiones presenta mayor capacidad predictiva.[12]

[11] Siendo precisos, bajo nuestra definición de función de pérdida discreta g, las variables aleatorias Z_i solo surgen como tales si los elementos del conjunto A definido en el apartado 2.2 son números reales.

[12] Dada la definición $ZZ = Z_2 - Z_1$, $bp > 0$ significa que el conjunto de previsiones 1 presenta mayor capacidad predictiva que 2, y viceversa.

b) La segunda definición de igualdad de capacidad predictiva que podemos utilizar en este contexto es más exigente que la anterior: "la probabilidad de que ZZ tome un valor negativo $-b_i$ es exactamente la misma que de tomar un valor positivo de la misma magnitud, $+b_i$", es decir:

$$H_0^{(2)} = p_i = p_{K-i+1},\ i = 1, 2, ..., q.$$

Es inmediato comprobar que $H_0^{(2)} \Rightarrow H_0^{(1)}$, pero no viceversa. La diferencia es obvia: por ejemplo, el conjunto 1 podría "ganar" a 2 pocas veces pero siempre por diferencia alta, mientras el modelo 2 podría ser preferible a 1 más veces pero con menor diferencia en tales casos. Según la definición $H_0^{(1)}$, podrían considerarse ambos conjuntos iguales en términos de bondad predictiva, mientras de acuerdo a $H_0^{(2)}$, son diferentes.

En este segundo planteamiento, la hipótesis alternativa será siempre bilateral: $H_1^{(2)} = \neg H_0^{(2)}$.

La última definición de hipótesis de igualdad predictiva lleva consigo dos problemas: por un lado, los contrastes que utilicen $H_0^{(2)}$ pueden ser menos potentes en la detección de diferencias de capacidad predictiva al no tener en cuenta los valores numéricos del vector b (no es lo mismo detectar cierta diferencia entre las probabilidades asociadas a los valores $-b_1$ y b_1 (los más pequeños) que entre $-b_q$ y b_q (los mayores)). Por otro lado, el hecho de que forzosamente la hipótesis alternativa deba ser bilateral supone cierta limitación en el contraste cuando existían a prioris sobre cuál es el modelo superior predictivamente.

Como veremos a continuación, el contraste Mult2 y su versión aproximada emplean las hipótesis $H_0^{(1)}$ y $H_1^{(1)}$ como hipótesis nula y alternativa, respectivamente, mientras RV-p utiliza $H_0^{(2)}$ y $H_1^{(2)}$.

Antes de pasar a exponer los contrastes, queremos hacer hincapié en el hecho de que ha sido fundamental establecer los contrastes sobre la muestra $(zz_1, ..., zz_T)$ y no sobre las dos muestras $(z_{1,1}, ..., z_{T,1})$ y $(z_{1,2}, ..., z_{T,2})$, para eludir el problema de la dependencia entre los dos conjuntos de previsiones, situación prácticamente inevitable en las comparaciones de capacidad predictiva en la práctica.

3.2.2. Contraste Mult2

La forma más natural de contrastar si $E(ZZ) = bp = 0$ sería a través de la media muestral $\overline{ZZ} = b\widehat{p} = T^{-1}bn$, donde n denota el vector de frecuencias $n = (n_1, ..., n_K)'$. Suponiendo conocida la distribución de $\overline{ZZ}$, que denotamos por $F_{\overline{ZZ}}$, compararíamos el valor observado del estadístico $\overline{ZZ}$ con el percentil apropiado de $F_{\overline{ZZ}}$ para un nivel de significación α, rechazando la hipótesis citada si el valor observado del estadístico es mayor que dicho percentil (o menor, según sea la región crítica del contraste). Pues bien, la distribución asintótica de $\overline{ZZ}$ se conoce y es estándar, como veremos en el apartado 3.2.3. Sin embargo, la distribución $F_{\overline{ZZ}}$ exacta, que querríamos aplicar en muestras finitas, no es conocida a priori. A continuación razonamos cómo obtener una estimación de ésta.

La distribución $F_{\overline{ZZ}}$ Como se afirmó en el primer punto del apartado 3.2.1, el vector de frecuencias $n = (n_1, n_2, ..., n_K)'$ sigue una distribución Multinomial de parámetros $p_1, p_2, ..., p_K$ y $T = n_1 + ... + n_K$. Pues bien, de este resultado se puede deducir de forma inmediata la probabilidad asociada a cualquiera de los valores posibles del estadístico $ZZ_n = bn = T\overline{ZZ}$ y, por tanto, deducir la distribución F_{ZZ_n} para muestras finitas. Obviamente, $F_{\overline{ZZ}}(w) = F_{ZZ_n}(Tw)$, así que cualquier test basado en el estadístico $\overline{ZZ}$ y su función de distribución $F_{\overline{ZZ}}$ es equivalente a otro usando ZZ_n y F_{ZZ_n}, por lo que nos vamos a centrar en definir ésta última función de distribución. El procedimiento es el siguiente:

(a) Se obtiene el conjunto de todos los posibles vectores de frecuencia asociados al tamaño muestral T. Dicho de manera formal, se trata de obtener el conjunto $\Psi = \{n = (n_1, n_2, ..., n_K)' | n_1 + n_2 + ... + n_K = T\}$, siendo n_i entero no negativo. Se ha diseñado un algoritmo apropiado para obtener todos estos vectores (véase Apéndice D, a final del documento).

(b) Cada uno de los vectores de frecuencia de Ψ se premultiplica por el vector b, obteniéndose el vector D_f, que contendrá todos los posibles valores x_i que puede tomar el estadístico $ZZ_n = bn$. Cada uno de ellos aparecerá repetido en D_f cierto número de veces $j(i)$, ya que, en general, habrá varios vectores de frecuencia que verifiquen $bn = x_i$). Sean $\{x_1, x_2, ..., x_L\}$ el conjunto de los valores distintos que aparecen en D_f, es decir, el soporte de la variable aleatoria ZZ_n.

(c) Para calcular el valor de la función de cuantía en un punto x_i, $P(ZZ_n = x_i)$, deben seleccionarse los vectores de frecuencia de Ψ que verifican $bn = x_i$, y calcular la probabilidad de que la muestra de longitud

T generara cualquiera de los vectores de frecuencia seleccionados. Dicha probabilidad viene determinada por la distribución Multinomial de parámetros $p_1, p_2, ..., p_K$, T. Formalmente:

– Se selecciona de Ψ el subconjunto $\Theta = \{n \in \Psi | ZZ_n = bn = x_i\}$.

– $P(ZZ_n = x_i) = P(n \in \Theta) = \sum_{n \in \Theta} P_M(n|p_1, ..., p_K, T)$, siendo $P_M(.)$ la función de cuantía de una distribución Multinomial, a saber:

$$P_M(n_1, n_2, ..., n_K|p_1, ..., p_K, T) = \prod_{i=1}^{K} (p_i)^{n_i} \frac{T!}{n_1! n_2! ... n_K!} \tag{5}$$

(d) Finalmente, se define la función de distribución: $F_{ZZ_n}(x_i) = \sum_{s=1}^{i} P(ZZ_n = x_s)$.

Dicha distribución de probabilidad depende no solo de los parámetros $p = (p_1, p_2, ..., p_K)'$ y T, sino también del vector b. La distribución en cuestión ni es estándar ni está tabulada, sino que debe construirse explícitamente para cada vector $(p', b', T)'$.

El contraste Proponemos un test para contrastar $H_0^{(1)}$, frente a la hipótesis alternativa $H_1^{(1)}$. Consistirá en (i) calcular el valor observado del estadístico ZZ_n, valor que denotaremos por $ZZ_n^0 = bn^0$, siendo n^0 el vector de frecuencias obtenido en la muestra, (ii) *estimar* la función de distribución F_{ZZ_n}, estimación que denotaremos por $\widehat{F}_{ZZ_n}$ y, finalmente, (iii) sea α el nivel de significación fijado por el usuario, aplicar la regla de decisión "se rechaza si solo si $-\left|ZZ_n^0\right| \leq \lambda_\beta$", siendo λ_β el valor tal que $\widehat{F}_{ZZ_n}(\lambda_\beta) = \beta$, donde $\beta = \alpha$ si la hipótesis es unilateral, mientras $\beta = \alpha/2$ si es bilateral.[13] En realidad, no es necesario estimar la función F_{ZZ_n} completa, sino que siempre bastará con estimar la cola inferior.

La estimación de F_{ZZ_n} se realiza siguiendo los pasos (a)-(d) del apartado precedente solo que usando estimaciones de los parámetros $p_1, p_2, ..., p_K$, que son desconocidos. Existen dos alternativas para estimar $p = (p_1, ..., p_K)'$: usando MV (Máxima Verosimilitud sin restricciones) ($\widehat{p}_i = \frac{n_i}{T}$) o MVR (Máxima Verosimilitud Restringida) bajo la restricción impuesta por la hipótesis $H_0^{(2)}$ del apartado 3.2.1.[14] Ésta última es la estimación de p que usaremos, porque es más acorde con la hipótesis nula $H_0^{(1)}$ y porque, además, es más eficiente, ya que estima la mitad de parámetros que MV, circunstancia ésta que tendrá valor en casos de longitud muestral pequeña. Volveremos a este asunto en el subapartado siguiente. El estimador MVR que emplearemos se denotará por $\overline{p}$ y su definición es:

$$\overline{p}_i = \overline{p}_{K-i+1} = \frac{n_i + n_{K-i+1}}{2T} = \frac{1}{2}\left(\widehat{p}_i + \widehat{p}_{K-i+1}\right), \text{ para } i = 1, ..., q = s-1; \ \overline{p}_s = \widehat{p}_s, \tag{6}$$

siendo s el subíndice correspondiente al valor 0 del vector b (es decir, $b_s = 0$). Usando (6), basta estimar los primeros q elementos de $\overline{p}$ (utilizando para el cálculo de $\overline{p}_i$ la información n_i y n_{K-i+1}), ya que los últimos q elementos tendrán la misma estimación que los primeros, mientras $\overline{p}_s = 1 - \overline{p}_1 - ... - \overline{p}_q$. Por lo tanto, tal y como afirmamos antes, se estimarán solo $q = \frac{K-1}{2}$ parámetros en vez de los $K-1$ parámetros que se estimarían por el método MV (todos menos $\widehat{p}_s$).

A modo de resumen, detallamos la secuencia de pasos para implementar el contraste:

(1) Dada la muestra $\{zz_t\}_{t=1}^T$, se calcula el valor observado del estadístico ZZ_n por $ZZ_n^0 = bn^0 = zz_1 + ... + zz_T$.

(2) Se estiman los puntos necesarios de la función F_{ZZ_n} usando los pasos (a)-(d) del apartado anterior, solo que, en el paso (c), se sustituye el vector paramétrico p de valor desconocido por su estimación MVR $\overline{p}$.

La simetría impuesta por construcción en el vector b (repásese el apartado 2.3) y en el estimador $\overline{p}$ garantizan que la función $\widehat{F}_{ZZ_n}$ sea simétrica en torno a $ZZ_n = 0$.[15] Debido a esta propiedad, podemos fijar

[13] En breve justificaremos la regla de decisión especificada.

[14] Lo deseable es utilizar un estimador de p que verifique la hipótesis nula $H_0^{(1)}$. Como ésta solo especifica una restricción muy general sobre el vector p y no permite inferir un mecanismo concreto de estimación, utilizamos la restricción $H_0^{(2)}$ que sí lo permite y garantiza a la vez que el vector estimado cumpla $H_0^{(1)}$.

[15] Esta afirmación es cierta para $\widehat{F}_{ZZ_n}$, pero no para la verdadera función F_{ZZ_n}, ya que el verdadero vector p podría no ser simétrico –es decir, $p_i \neq p_{K-i+1}$ para algún i–, incluso aunque la hipótesis nula $H_0^{(1)}$ fuera cierta, y mucho menos, si fuera falsa.

la cola inferior siempre como región crítica del contraste, y utilizar la regla de decisión arriba especificada. Por ello, los puntos de $\widehat{F}_{ZZ_n}$ que deben calcularse son solamente los correspondientes a los primeros puntos del soporte $x_1, x_2, ..., x_l = \lambda_\beta$, siendo $\widehat{F}_{ZZ_n}(x_l) = \beta$.

(3) Aplicar la regla de decisión del contraste que expusimos antes: "se rechaza si solo si $-\left|ZZ_n^0\right| \leq \lambda_\beta$", siendo λ_β el valor para que verifica $\widehat{F}_{ZZ_n}(\lambda_\beta) = \beta$, siendo $\beta = \alpha$ si la hipótesis es unilateral, y $\beta = \alpha/2$ si es bilateral.[16]

Evaluación del test: calidad de la aproximación a la distribución exacta Obviamente, si la distribución exacta $F_{\overline{ZZ}}$ fuera conocida, el test sería también exacto en tamaño, en muestras finitas.[17] Sin embargo, el contraste procede estimando la distribución de probabilidad, a través del estimador:

$$\widehat{P}(\overline{ZZ} = w) = \sum_{n \in \Theta} P_M(n|\overline{p}, T), \text{ donde } P_M \text{ se define en (5)}, \tag{7}$$

siendo Θ el conjunto de vectores de frecuencia n tales que $bn = Tw$, y $\overline{p}$ el estimador MVR definido en (6).

Es decir, las probabilidades de la distribución de $\overline{ZZ}$ se estiman a través de la expresión exacta de éstas, solo que sustituyendo el vector desconocido p por su estimador $\overline{p}$.

Por tanto, la calidad de la aproximación a la distribución teórica que usa el contraste está muy relacionada con las propiedades del estimador $\overline{p}$. Bajo la hipótesis nula y suponiendo que ésta se verifica porque $p_i = p_{K-i+1}$, para $i = 1, 2, ..., s-1$ (condición suficiente, pero no necesaria para $bp = 0$), es obvio que $\overline{p}$ es un estimador insesgado y consistente para p, propiedades que, por supuesto, también cumple el estimador MV $\widehat{p}$. Sin embargo, como ya se mencionó arriba, $\overline{p}$ es más eficiente que $\widehat{p}$, ya que es un estimador restringido, limitando, por tanto, el espacio de búsqueda, lo que reduce su varianza. A este respecto, se tienen los siguientes resultados, válidos bajo el supuesto $p_i = p_{K-i+1}$:

a) $traza(E[(\overline{p}-p)(\overline{p}-p)']) = \sum_{i=1}^{K} V(\overline{p}_i) = \frac{1}{2T}(1+p_s) - \frac{1}{T}\sum_{i=1}^{K} p_i^2$. Este resultado se deduce fácilmente del supuesto citado, de la definición del estimador MVR $\overline{p}$ ($\overline{p}_i = \frac{1}{2}\widehat{p}_i + \frac{1}{2}\widehat{p}_{K-i+1}$) y de las conocidas expresiones de las varianzas y covarianzas teóricas de los estimadores MV $\widehat{p}$ ($V(\widehat{p}_i) = \frac{1}{T}p_i(1-p_i)$, $E(\widehat{p}_i - p_i)(\widehat{p}_j - p_j) = -\frac{1}{T}p_i p_j$).[18]

b) $\sum_{i=1}^{K} V(\widehat{p}_i) - \sum_{i=1}^{K} V(\overline{p}_i) = \frac{1}{2T}(1-p_s) > 0$. Esto es inmediato a partir de a) y de la expresión de la varianza de $\widehat{p}_i$.

Es decir, la "varianza agregada" de $\overline{p}$, medida por $\sum_{i=1}^{K} V(\overline{p}_i)$, es siempre menor que la asociada a $\widehat{p}$, y dicha diferencia disminuye conforme mayor es p_s, es decir, conforme mayor es la probabilidad asociada al suceso $\{ZZ = 0\}$. En relación con todo este asunto, podemos concluir los siguientes importantes resultados:

1. Eligiendo $\overline{p}$ como estimador de p, en vez de $\widehat{p}$, es razonable pensar que se logrará mayor exactitud en el tamaño del test, al ser más precisa la estimación de $F_{\overline{ZZ}}$. Esta ventaja probablemente se atenúa conforme mayor es p_s.

2. Por otro lado, anticiparemos un resultado fundamental que hemos observado empíricamente: la aproximación $\widehat{F}_{\overline{ZZ}}$ funciona peor conforme mayor es p_s, pese a que un incremento de p_s no implica necesariamente un aumento de la "varianza agregada" de $\overline{p}$. Es decir, aunque no haya mayor imprecisión en la estimación $\overline{p}$ cuando p_s es muy elevado, la varianza del estimador (7) sí parece aumentar, según mostrarán los resultados de nuestras simulaciones. No obstante, justificar analíticamente esta afirmación es un problema que consideramos intratable.

[16] En la práctica, la regla de decisión no será exactamente ésta, puesto que no existirá, salvo por casualidad, ningún punto de la función de distribución estimada para el que se verifique $\widehat{F}_{ZZ_n}(\lambda_\beta) = \beta$. En el apartado segundo de 3.2.5 se explica detalladamente esta cuestión. De momento, solo se pretende presentar conceptualmente el test y, a estos efectos, no vale la pena incidir en este asunto, que, sin embargo, tiene relevancia en la práctica.

[17] En realidad, esta afirmación es correcta si el test se implementa en su versión aleatorizada. El asunto de la aleatorización se trata en el segundo apartado de 3.2.5.

[18] $V(\overline{p}_i) = V(\frac{1}{2}\widehat{p}_i + \frac{1}{2}\widehat{p}_{i'}) = \frac{1}{4T}(p_i(1-p_i) + p_{i'}(1-p_{i'}) - 2p_i p_{i'})$, donde p_{K-i+1} se ha denotado por $p_{i'}$. Usando $p_i = p_{i'}$ se tiene que $V(\overline{p}_i) = \frac{1}{2T}(p_i - 2p_i^2)$.

Por tanto, $\sum V(\overline{p}_i) = \sum_{i \neq s} \frac{1}{2T}(p_i - 2p_i^2) + \frac{1}{T}(p_s - p_s^2) = \frac{1}{2T}\sum_{\forall i} p_i + \frac{1}{2T}p_s - \frac{1}{T}\sum_{\forall i} p_i^2 = \frac{1}{2T}(1+p_s) - \frac{1}{T}\sum_{\forall i} p_i^2$, q.e.d.

Algunos problemas de implementación del test Mult2 conlleva algunos problemas técnicos para su implementación:

a) Se requiere independencia de las variables zz_t para que la derivación del contraste sea correcta. En caso de incumplimiento de dicho supuesto, se aplicará el procedimiento que se explica en el cuarto apartado de 3.2.5.

b) El cálculo del test es costoso computacionalmente. Es fácilmente demostrable que la cardinalidad del conjunto Ψ es $CR(K,T) = C(T+K-1,T) = \frac{T+K-1!}{T!K-1!} = \frac{(T+K-1)(T+K-2)...(T+1)}{K-1!}$, donde CR y C designan combinaciones con y sin repetición, respectivamente. De acuerdo a esto, cualquier algoritmo para la implementación presenta un coste computacional $O(T^{K-1})$,[19] lo que puede convertir el test en poco operativo si la muestra es larga o/y si el número de valores distintos K que puede tomar la función f que compara las pérdidas es relativamente elevado. Para dichas situaciones, conviene sustituir el test Mult2 por su versión asintótica, que presentamos en el siguiente apartado.

En el Apéndice D se ofrece un breve análisis computacional de Mult2, a nivel teórico y empírico, que permitirán al lector hacerse una idea del tiempo que puede requerir la ejecución del contraste. En general, podría decirse que Mult2 puede ejecutarse en tiempos muy razonables si $T \leq 50$ cuando $K = 5$, mientras si $K = 7$, solo es aconsejable para longitudes muestrales $T \leq 25$. Las cifras anteriores se corresponden al caso en que la serie zz_t no presente autocorrelación, situación que, normalmente, ocurrirá si el horizonte de previsión es uno. En cambio, si existe autocorrelación de orden r (en principio, cuando el horizonte de previsión sea $r+1$), el test sigue siendo admisible en términos de coste computacional para longitudes muestrales $r+1$ veces mayores que las citadas anteriormente. Esto es debido a que, al aparecer autocorrelación de orden r, se aplicará el procedimiento explicado en el apartado cuarto de 3.2.5, que dividirá la muestra de longitud T en $r+1$ submuestras de longitud $T/(r+1)$, y el contraste se aplicará sobre las submuestras.

c) En teoría, el test no podría aplicarse si $\exists i$ tal que $n_i + n_{K-i+1} = 0$, ya que $\overline{p}_i = 0$ y no se podrían calcular las probabilidades (5) ni, en consecuencia, estimar $F_{\overline{ZZ}}$. La solución a este problema se abordará en el primer apartado de 3.2.5. La corrección propuesta es muy sencilla y, como veremos, prácticamente no afecta las propiedades de tamaño y potencia del test.

3.2.3. Versión aproximada de Mult2 (Contraste Mult2-aprx)

Mult2 utiliza una estimación de la función de distribución exacta del estadístico $\overline{ZZ}$. Sin embargo, si T es grande, el test se convierte en poco operativo, por el coste de estimar $F_{\overline{ZZ}}$. Pero, precisamente cuando T es grande, el Teorema Central del Límite permite deducir la distribución asintótica del estadístico $\overline{ZZ}$, que podremos usar como aproximación de la verdadera $F_{\overline{ZZ}}$ en muestras suficientemente largas. Así, todo el procedimiento expuesto para Mult2 puede sustituirse por el cálculo del estadístico $\overline{ZZ}$ y la comparación de éste con el percentil apropiado de una distribución Normal. En concreto:

Utilizando el resultado sobre la convergencia en ley del estimador MV del vector p en una distribución Multinomial, se tiene que $\sqrt{T}(\widehat{p}-p) \xrightarrow{L} N(0,V_p)$, siendo $V_p = \Omega - pp'$ y Ω una matriz diagonal $K \times K$ con los elementos de p en la diagonal. Por lo tanto, puede contrastarse la hipótesis $H_0^{(1)}$ frente a $H_1^{(1)}$ a través del estadístico $\sqrt{T}b\widehat{p}\widehat{W}_p^{-1/2}$ y de su distribución límite $N(0,1)$, siempre que $\widehat{W}_p$ represente una estimación *consistente* de $W_p = bV_pb'$.

Existen dos estimadores consistentes de W_p. El primero lo denotaremos por $\widehat{W}_{\widehat{p}}$, y consiste en utilizar la expresión W_p pero sustituyendo p por su estimación de máxima verosimilitud (MV), es decir, $\widehat{p}$ ($\widehat{p}_i = \frac{n_i}{T}$). Dicho estimador es consistente siempre, se cumpla o no la hipótesis nula, ya que $\widehat{p}$ es un estimador consistente de p. Como veremos a continuación, cuando las pérdidas no estén autocorreladas, el contraste propuesto y el test DM son el mismo si el primero se implementara usando $\widehat{W}_{\widehat{p}}$, y siempre que esté utilizando la diferencia $ZZ = f(Z_1,Z_2) = Z_2 - Z_1$ como función de comparación de pérdidas. El segundo estimador lo denotamos por $\widehat{W}_{\overline{p}}$ y consiste es sustituir p en la expresión W_p por su estimador MV restringido a la hipótesis $H_0^{(2)}$, que simbolizamos por $\overline{p}$ y cuya formulación se presentó en (6). Como se demostrará en el Apéndice C, $\widehat{W}_{\overline{p}}$ es siempre un estimador consistente de W_p bajo la hipótesis $H_0^{(1)}$, pese a que $\overline{p}$ no lo es para p, en general (es consistente siempre bajo $H_0^{(2)}$, pero puede no serlo bajo $H_0^{(1)}$). Por coherencia

[19] En el Apéndice 4 se presenta un análisis computacional del test. Allí puede consultarse la definición formal de la expresión $O(f(n))$ para una función $f(n)$.

con lo que hacemos en Mult2, donde se emplea $\overline{p}$ como estimador del vector p, que es, al fin y al cabo, el test para el que pretendíamos obtener su versión asintótica, y por razones de eficiencia en la estimación, que luego trataremos, emplearemos $\widehat{W}_{\overline{p}}$ en nuestro contraste, que llamaremos Mult2-aprx. Así por tanto, Mult2-aprx queda finalmente definido por:

$$M2_{apx} = \sqrt{T} b \widehat{p} \widehat{W}_{\overline{p}}^{-1/2} \underset{H_0^{(1)}}{\overset{L}{\rightarrow}} N(0,1), \tag{8}$$

$$\widehat{W}_{\overline{p}} = b V_{p|\overline{p}} b'.$$

La región crítica del contraste será una de las colas de la distribución $N(0,1)$ con masa de probabilidad α (si $H_1 = H_1^{(1u)}$), o bien ambas con masa $\alpha/2$ en cada una (si $H_1 = H_1^{(1d)}$) según se dispongan de aprioris sobre el modelo mejor en términos predictivos.

Las únicas limitaciones de implementación que tiene el test son las siguientes:

a) Igual que para Signos, Wilcoxon, Mult2 y para el contraste que se presentará en la siguiente sección, la derivación de la distribución del estadístico $M2_{apx}$ presupone aleatoriedad en la muestra $\{zz_t\}_{t=1}^{T}$, por lo que, en caso de incumplirse el supuesto, el test Mult2-aprx debe aplicarse utilizando el procedimiento del cuarto apartado de 3.2.5.

b) En teoría, el test no podría llevarse a cabo si $\widehat{W}_{\overline{p}} = 0$. En el Apéndice C se demuestra que el estimador $\widehat{W}_{\overline{p}}$ puede escribirse como $\widehat{W}_{\overline{p}} = \sum_{i=1}^{K} b_i^2 \widehat{p}_i \geq 0$ y, por lo tanto, el único caso en el que $\widehat{W}_{\overline{p}} = 0$ es aquel en que las frecuencias absolutas son $n_s = T$ y $n_i = 0 \ \forall i \neq s = \frac{K+1}{2}$ (recuérdese que $b_s = 0$). Es decir, cuando en *todos* los periodos de la muestra, las dos previsiones a comparar obtuvieron la misma penalización (formalmente, $z_{t,1} = z_{t,2}$, $f(z_{t,1}, z_{t,2}) = zz_t = 0 \ \forall t$). Por analogía con la forma en que se abordaron las situaciones de frecuencias nulas en Mult2, se aplicará el mismo procedimiento que en dicho caso, procedimiento sencillo que se explica en el primer apartado de 3.2.5. Como se verá allí, el resultado es que, cuando se dé esta situación, el test Mult2-aprx no rechazará la hipótesis nula, tal y como hubiera sido deseable, ya que es el caso más claro posible de igualdad de capacidad predictiva entre los dos conjuntos de previsión.

Relación entre DM y Mult2-aprx Como ya se ha mencionado anteriormente, Mult2-aprx y DM son contrastes parecidos, así que es razonable pensar que podemos obtener algunos resultados teóricos sobre dichas similitudes, antes de ponernos a explorar sus propiedades estadísticas en la sección posterior vía simulación. Obtendremos resultados teóricos sobre la relación entre los dos tests, algunos de ellos exactos y otros con validez asintótica. Aquellas propiedades que no puedan ser establecidas analíticamente, se tratarán de obtener a través de ejercicios de simulación en esta misma sección, siempre que éstas sean de carácter asintótico. Las de muestras finitas son objeto de estudio en la sección 4, junto con el resto de contrastes.

Sea el siguiente conjunto de condiciones (a)-(d), que denotamos por [C]:

(a) La varianza $2\pi f_d(0)$ del estadístico $\sqrt{T}\overline{d}$ de DM se estima mediante la suma de las autocovarianzas muestrales del diferencial d_t de órdenes $\tau = -(h-1), ..., h-1$, siendo h el horizonte de previsión;

(b) El horizonte de previsión h es igual a uno. Por tanto, usando (a), se deduce que la estimación de la varianza $2\pi f_d(0)$ es, simplemente, $\frac{1}{T}\sum_{t=1}^{T}(d_t - \overline{d})^2$;

(c) La función de pérdida $g(y_t, v_{t,i})$ empleada en el cálculo de DM es la misma función de pérdida discreta que en Mult2-aprx.

(d) La función de comparación de pérdidas f empleada en Mult2-aprx es $f(z_{t,1}, z_{t,2}) = zz_t = z_{t,2} - z_{t,1}$. Es decir, zz_t es el diferencial d_t del test DM.

Denotemos por S_1^D el estadístico del test DM si se aplicara bajo el conjunto de condiciones [C].

I. Resultados teóricos sobre la relación entre DM y Mult2-aprx

Se tienen los siguientes resultados, demostrados en el Apéndice C:

1. Estimadores de W_p considerados para Mult2-aprx:

[Prop 1] $\widehat{W}_{\widehat{p}} = \sum_{i=1}^{K} b_i^2 \widehat{p}_i - (b\widehat{p})^2$; es un estimador consistente de W_p.

[Prop 2a] $\widehat{W}_{\overline{p}} = \sum_{i=1}^{K} b_i^2 \widehat{p}_i$; es un estimador consistente de W_p si $H_0^{(1)}$ es cierta.

[Prop 2b] $V(\widehat{W}_{\overline{p}}) = T^{-1} b^{(2)} V_p b^{(2)\prime}$, donde $b^{(2)}$ es un vector $1 \times K$ cuyo elemento iésimo es $b_i^{(2)} = b_i^2$.

Las propiedades a continuación se cumplen si y solo si se verifica el conjunto de condiciones [C]:

2. Relación exacta entre $M2_{apx}$ y S_1^D:

[Prop 3] $S_1^D = \varphi(M2_{apx}) = \frac{M2_{apx}}{\sqrt{1 - \frac{(M2_{apx})^2}{T}}}$.

Otra forma de escribir la relación, útil en algunos casos, es:

[Prop 4] $\frac{M2_{apx}}{S_1^D} = \left(1 - \frac{(b\widehat{p})^2}{\sum_{i=1}^{K} b_i^2 \widehat{p}_i} \right)^{1/2}$.

3. Corolarios deducidos a partir de [Prop 4]:

3.a. [Prop 5] "Si Mult2-aprx usara $\widehat{W}_{\widehat{p}}$ como estimación de W_p, en vez de $\widehat{W}_{\overline{p}}$, entonces $M2_{apx}$ sería exactamente igual a S_1^D".

3.b. [Prop 6] "Si la hipótesis nula $H_0^{(1)}$ es cierta, $\frac{M2_{apx}}{S_1^D} \xrightarrow{p} 1$. En el resto de casos, se cumple $\frac{M2_{apx}}{S_1^D} \xrightarrow{p} \left(1 - \frac{(bp)^2}{\sum_{i=1}^{K} b_i^2 p_i} \right)^{1/2}$".

4. Corolarios deducidos a partir de [Prop 3]:

4.a. [Prop 7] "El estadístico S_1^D está definido en todos los puntos en los que lo está $M2_{apx}$ (por su parte, $M2_{apx}$ está definido para cualquier muestra salvo aquella que genere que la frecuencia relativa central $\widehat{p}_s$ sea igual a uno), salvo en los puntos $M2_{apx} = -\sqrt{T}$ y $M2_{apx} = +\sqrt{T}$ (que se corresponden con muestras que generaron alguna frecuencia relativa no central $\widehat{p}_i$ igual a uno, para algún $i \neq s$)".

Por tanto, siempre que $M2_{apx}$ no esté definido, tampoco lo estará S_1^D, pero lo contrario no es cierto. Como se verá en el primer apartado de 3.2.5 (al referirnos a Mult2-aprx) y en el penúltimo de la misma subsección (al referirnos a DM), la implementación de los contrastes en estos casos se soluciona de forma simple en la práctica, garantizando que la decisión del test sea la correcta: no rechazar H_0 si la muestra generó $\widehat{p}_s = 1$, y rechazar si generó $\widehat{p}_i = 1$, para cualquier $i \neq s$.

4.b. [Prop 8] "Se verifica $\left|S_1^D\right| > |M2_{apx}|$ en todos los puntos del soporte de $M2_{apx}$, salvo en el punto $M2_{apx} = 0$ (para el que $S_1^D = 0$)".

4.c. [Prop 9] "La función φ es monótona (creciente), y, por tanto, $P(S_1^D = \varphi(x)) = P(M2_{apx} = x)$, para cualquier punto $x \in (-\sqrt{T}, +\sqrt{T})$ del soporte de la variable aleatoria $M2_{apx}$ (siendo los extremos de dicho soporte (incluidos en él), precisamente, $-\sqrt{T}$ y $+\sqrt{T}$)". Por lo tanto, las funciones de distribución de ambos estadísticos son la misma, solo que para soportes diferentes. Esta propiedad es de gran relevancia para entender los resultados, teóricos y obtenidos por simulación, sobre la comparación entre ambos tests en términos de su potencia, siempre que se mida por el concepto de SAP.

II. Conclusiones sobre la comparación entre Mult2-aprx y DM en tamaño y potencia

En la mayor parte de aplicaciones de DM para funciones de pérdida discretas y en todas las simulaciones de este capítulo de la tesis y del siguiente, las condiciones (a), (c) y (d) se verificarán. Por ello, los resultados [Prop 1]-[Prop 9] nos permiten obtener conclusiones sobre la relación entre DM y Mult2-aprx, siempre que nos restrinjamos a horizonte de previsión uno. A continuación exponemos esas conclusiones teóricas, y planteamos otras cuestiones sobre la relación entre los contrastes, que deben resolverse por simulación. De ellas, resolveremos en este mismo apartado las que se circunscriban al caso asintótico, posponiendo la evaluación y comparación de propiedades para muestras finitas para la sección 4, con el resto de tests:

1. Horizonte de previsión uno: De las propiedades [Prop 6], [Prop 8] y [Prop 9] enunciadas arriba, se deducen las siguientes afirmaciones, *que se restringen al caso* $h = 1$ (y a la versión discreta de DM, se entiende):

(a) La probabilidad de rechazar la hipótesis nula es siempre mayor usando DM que con Mult2-aprx. Esta conclusión se obtiene de [Prop 8] de forma inmediata. La causa es que $\widehat{W}_{\overline{p}} \geq \widehat{W}_{\widehat{p}}$ ya que, como afirma [Prop 5], S_1^D es igual a $M2_{apx}$ salvo que utiliza como varianza $\widehat{W}_{\widehat{p}}$, mientras $M2_{apx}$ usa $\widehat{W}_{\overline{p}}$.

(b) Respecto al *tamaño* de los contrastes:

– Debido a [Prop 6], DM y Mult2-aprx son *asintóticamente equivalentes* bajo $H_0^{(1)}$ ([Prop 6]), por lo que el tamaño de ambos tests debe coincidir en muestras largas.

– En muestras finitas, sabemos, en primer lugar, que el tamaño del test Mult2-aprx debe ser menor que el de DM, debido a la afirmación (a). Pero, aparte de esto, tenemos razones para pensar que *Mult2-aprx será más exacto en tamaño que DM*, en el caso $h = 1$. La justificación de esta afirmación es la siguiente: en el caso $h = 1$, una versión equivalente de estos dos tests es utilizar el estadístico $b\widehat{p}$ como estadístico de contraste y la distribución asintótica $N(0, T^{-1}W_p)$ como aproximación a la distribución exacta de $b\widehat{p}$. Pero la distribución asintótica se estima, porque p es desconocido. Por lo tanto, cuanto más eficiente sea dicha estimación, más precisa será la aproximación a la distribución asintótica (que, como veremos después, es una buena aproximación a la verdadera distribución de $b\widehat{p}$), y el tamaño del test será más exacto. Pero, como afirma [Prop 5], la diferencia entre ambos tests es precisamente que Mult2-aprx usa $\widehat{W}_{\overline{p}}$ como estimador de W_p, mientras DM usa $\widehat{W}_{\widehat{p}}$. Ambos son estimadores insesgados y consistentes bajo $H_0^{(1)}$, pero, según [Prop 1] y [Prop 2], se tiene que $\widehat{W}_{\widehat{p}} = \widehat{W}_{\overline{p}} - (b\widehat{p})^2$, por lo que la varianza de $\widehat{W}_{\widehat{p}}$ será mayor que la de $\widehat{W}_{\overline{p}}$. Para cerciorarnos del efecto de esta discrepancia sobre el tamaño de los tests, y anticipando lo que se hará en los ejercicios de simulación de la sección 4, llevaremos a cabo esos mismos ejercicios ahora para más longitudes muestrales, comprobando cómo de rápido se ajustan al teórico los tamaños empíricos de ambos contrastes. Denotaremos este asunto como "cuestión A".

(c) Respecto a la *potencia* de los contrastes: Si $H_0^{(1)}$ no se cumple, es decir, si $bp \neq 0$, los tests DM y Mult2-aprx no son asintóticamente equivalentes (su ratio no converge a uno, sino al valor indicado en [Prop 6]). Sin embargo, si la medición de la potencia se realiza a través del concepto de SAP, es decir, ajustando la potencia según tamaño, *la potencia de ambos contrastes será exactamente la misma*, en muestras finitas y asintóticamente, en virtud de [Prop 9]. Para entender tal afirmación, léase el último punto del Apéndice C (si el lector no está familiarizado con la medida SAP, también es conveniente leer el apartado sobre ella del Apéndice A). Las estimaciones de SAP obtenidas en todos los experimentos realizados en este capítulo, tanto en la sección actual como en la sección 4, verifican el resultado en cuestión.

2. Horizonte de previsión mayor que uno: No disponemos de resultados teóricos que ayuden a concluir sobre la similitud o diferencia respecto al comportamiento asintótico entre el test Mult2-aprx y la versión discreta de DM en el caso $h > 1$. Véase que, para horizontes de previsión superiores a uno, las propiedades teóricas obtenidas anteriormente no aplican, ya que:

– La estimación de la varianza utilizada en el cálculo del estadístico de contraste de DM incluye autocovarianzas de orden superior a cero. Es decir, no se verifica la condición (b) de [C].

– La implementación de Mult2-aprx requiere el uso del procedimiento de la cota de Bonferroni.

En este caso, el análisis comparativo solo será posible a través de simulaciones, que presentamos a continuación. Denotamos ésta como "cuestión B". La comparativa en muestras finitas es objeto de la sección 4.

3. Conclusiones empíricas a las dos cuestiones pendientes: Se acaban de citar dos situaciones (denotadas por A y B) para las que sería deseable tener resultados de experimentos de simulación, de cara a establecer relaciones entre los contrastes DM y Mult2-aprx, en términos de propiedades estadísticas de tamaño y potencia. El análisis de Monte Carlo de la sección 4 obtiene conclusiones sobre las citadas propiedades para todos los tests que se evalúan en el capítulo (entre ellos, DM y Mult2-aprx) pero en muestras finitas. Para resolver las dudas A y B en términos asintóticos, se han realizado los mismos experimentos de simulación que los presentados en la próxima sección 4 pero solo para DM y Mult2-aprx, y usando muestras largas. Los resultados se exponen en la Tabla 17, en el Apéndice C. Las conclusiones son las siguientes:

a) El test Mult2-aprx converge al tamaño correcto más rápidamente que DM si $h = 1$ (cuestión A). Aproximadamente, y salvo en algunos casos excepcionales, Mult2-aprx se ajusta al tamaño teórico desde $T = 24$, mientras DM lo hace a partir de $T = 64$. La afirmación efectuada arriba sobre la diferencia de varianza entre los estimadores $\widehat{W}_{\widehat{p}}$ y $\widehat{W}_{\overline{p}}$ (más eficiente éste último) parece confirmarse. No obstante, el sesgo de DM en longitudes muestrales en el intervalo $[24, 64)$ es pequeño.

b) Cuando $h = 2$, obtenemos los siguientes resultados (cuestión B): (i) las diferencias en convergencia al tamaño correcto siguen aproximadamente el mismo patrón que en el caso $h = 1$, aunque las longitudes muestrales para las que se produce la convergencia son mayores, en ambos tests; (ii) las diferencias en el sesgo en tamaño de los tests hasta alcanzar dicha convergencia son analizadas en la sección 4, en los experimentos para muestras finitas, aunque se pueden observar también en la Tabla 17: el sesgo es considerablemente mayor en DM que en Mult2-aprx; (iii) en muestras de longitud intermedia, entre 24 y 128 datos, la SAP del contraste DM es mayor que la de Mult2-aprx. Ambas se igualan a partir de $T = 128$, en un nivel de potencia aproximadamente igual a uno. La diferencia de potencia en muestras muy cortas ($T < 24$) es favorable a Mult2-aprx, pero esto se estudiará en el análisis de la sección 4.

c) Como ya se había adelantado, los resultados de estos ejercicios confirman que la SAP es idéntica para ambos tests, cuando $h = 1$.

Diferencias teóricas entre Mult2 y Mult2-aprx. Calidad de la aproximación La distribución exacta del estadístico $b\widehat{p}$ es $F_{\overline{ZZ}}$, distribución introducida en el primer punto del apartado 3.2.2, cuando se presentó el test Mult2. Éste se basa en aproximar dicha distribución por una estimación, que consiste simplemente en calcular las probabilidades según la expresión correcta (5), solo que estimando el vector paramétrico desconocido p a través del estimador $\overline{p}$. El test Mult2-aprx, tal y como se ha mencionado anteriormente, utiliza implícitamente la distribución asintótica de $b\widehat{p}$, que es $N(0, T^{-1}W_p)$. Por lo tanto, la diferencia de propiedades entre Mult2 y Mult2-aprx solo depende de la calidad de la aproximación que hace cada uno. Las discrepancias son:

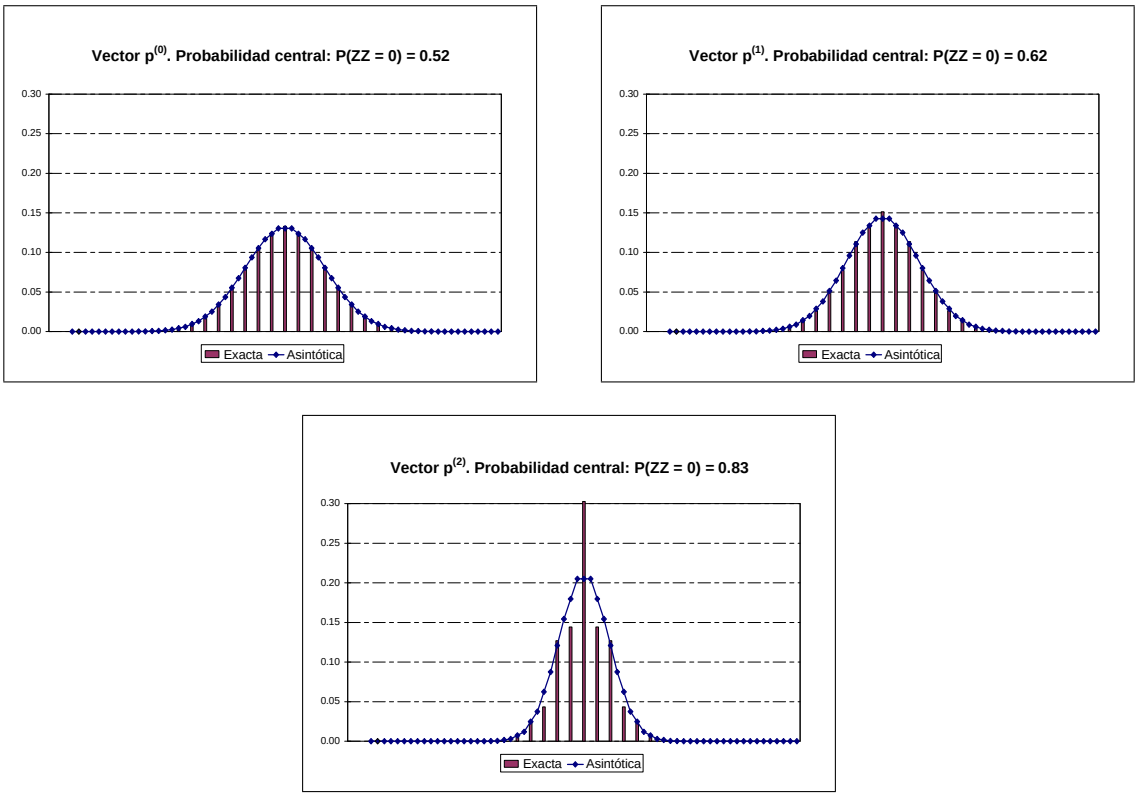
a) Mult2 parte con la ventaja de que trata de estimar la distribución exacta, no una aproximación a ella. La pregunta crucial es si la distribución asintótica $N(0, T^{-1}W_p)$ es muy diferente de la verdadera $F_{\overline{ZZ}}$ en tamaños muestrales pequeños. Obviamente, esto depende de los vectores b y p. Para indagar sobre este asunto, hemos generado ambas distribuciones teóricas para longitud muestral $T = 8$ bajo las mismas condiciones que se dan en los tres ejercicios de tamaño de los experimentos de la sección 4 para el caso $h = 1$,[20] es decir, usando los siguientes vectores: $b = (-2, -1, 0, +1, +2)$ en todos los casos, mientras los valores numéricos de p son $p^{(0)} = (0{,}108, 0{,}133, 0{,}518, 0{,}133, 0{,}108)'$, $p^{(1)} = (0{,}093, 0{,}096, 0{,}622, 0{,}096, 0{,}093)'$ y $p^{(2)} = (0{,}043, 0{,}043, 0{,}828, 0{,}043, 0{,}043)'$. Las distribuciones se presentan en la Figura 1. Como puede comprobarse, la aproximación asintótica es muy certera, salvo en el tercer caso, cuando $p_s = P(ZZ = 0)$ es muy elevado.

b) Por otro lado, Mult2 debe estimar cada probabilidad $P(\overline{ZZ} = b\widehat{p} = w)$ de la distribución exacta de $\overline{ZZ} = b\widehat{p}$. Si denotamos dicha probabilidad por $\zeta_w(p)$, el estimador utilizado por Mult2 es $\zeta_w(\overline{p})$, tal y como mostramos en (7). El estimador $\overline{p}$ es insesgado y consistente bajo $H_0^{(2)}$, y su "varianza agregada" (medida por la traza de su matriz de varianzas y covarianzas) se presentó en el tercer punto de 3.2.2. Por su parte, Mult2-aprx estima W_p a través de $\widehat{W}_{\overline{p}}$, es decir, estima $\sum_{i=1}^{K} b_i^2 p_i - (bp)^2$ por $\sum_{i=1}^{K} b_i^2 \widehat{p}_i$. Dicho estimador es también insesgado y consistente bajo $H_0^{(1)}$ (y bajo $H_0^{(2)}$, por supuesto), y su varianza se presentó en [Prop 2b]. A priori, no sabemos cuál de las dos estimaciones genera peor aproximación respecto a la distribución que trata de estimar, y ésta es la segunda fuente de diferencias entre ambos tests.

Los efectos de ambas discrepancias se tendrán que evaluar empíricamente, en los ejercicios de la sección 4. No obstante, la gran similitud que existe entre la distribución exacta $F_{\overline{ZZ}}$ y la asintótica $N(0, T^{-1}W_p)$ usada por Mult2-aprx, salvo cuando p_s es elevado, garantiza que emplear Mult2-aprx en vez de Mult2 puede ser en la práctica una forma de proceder muy razonable, si se desea evitar el coste computacional de la implementación de Mult2.

[20] En concreto, son las funciones de masa las que se dibujan. La correspondiente a la distribución exacta no requiere más explicación. Respecto a la asociada a la $N(0, T^{-1}W_p)$, la función se genera así: para cada dos puntos x_0 y x_1 del soporte de la distribución exacta, se presenta el punto $(x', A(x'))$, siendo $x' = \frac{x_0+x_1}{2}$ y $A(x') = \Phi_W(x_1) - \Phi_W(x_0)$, donde $\Phi_W(.)$ representa la función de distribución de una $N(0, T^{-1}W_p)$.

Figura 1. Distribuciones exacta y asintótica del estadístico $b\widehat{p}$, según valor del vector p. Valores utilizados: $p^{(0)}$, $p^{(1)}$ y $p^{(2)}$ (Experimentos de Tamaño). $b = (-2, -1, 0, +1, +2)$; $T = 8$



3.2.4. Contraste RV-p

En el contexto en que nos manejamos, donde la función de verosimilitud asociada a las frecuencias $n_1, n_2, ..., n_K$ es conocida, puede aplicarse la versión asintótica de un test paramétrico de Razón de Verosimilitudes para contrastar $H_0^{(2)}$ frente a $H_1^{(2)} = \neg H_0^{(2)}$ de forma sencilla.[21] Sean $\widehat{p}$ y $\overline{p}$ los estimadores MV bajo $H_1^{(2)}$ y $H_0^{(2)}$, respectivamente, donde $\widehat{p}_i = \frac{n_i}{T}$ y $\overline{p}$ se definió en (6). Basta utilizar la expresión (5) para representar la función de verosimilitud asociada a una distribución Multinomial y simplificar, para obtener:

$$rv_p = 2\sum_{i=1}^{K} n_i \left(\ln \widehat{p}_i - \ln \overline{p}_i\right) \underset{H_0^{(2)}}{\overset{L}{\rightarrow}} \chi_q^2, \tag{9}$$

siendo $q = \frac{K-1}{2}$.

El contraste es siempre bilateral y la región crítica es la cola superior de masa de probabilidad α de la distribución χ_q^2.

El contraste RV-p es muy sencillo de implementar pero presenta algunas limitaciones importantes:

a) Igual que Mult2-aprx, RV-p es de validez asintótica. Además, su derivación requiere aleatoriedad en la secuencia $\{zz_t\}_{t=1}^{T}$. En caso de incumplimiento de dicho supuesto, se aplicará el procedimiento explicado en el cuarto apartado de 3.2.5.

[21] En el caso que nos ocupa, el estadístico de razón de verosimilitudes Λ no tiene distribución conocida, por lo que solo es posible usar la distribución límite de $-2 \lg \Lambda$, es decir, la versión asintótica del test.

b) El rechazo de la hipótesis nula en favor de la alternativa no permite determinar directamente el conjunto de previsiones preferible.[22]

c) En teoría, el test no podría aplicarse si $\exists i$ tal que $n_i = 0$, ya que se tendría $\widehat{p}_i = 0$, imposiblitando el cálculo de la verosimilitud. La condición es bastante más restrictiva que la especificada para Mult2. Éste sería un problema extraordinariamente grave para el contraste en muestras cortas. Igual que se mencionó en el caso de Mult2, la solución que proponemos es muy sencilla y genera buenos resultados, y se expone en el primer apartado de 3.2.5.

3.2.5. Detalles de implementación de los tests

En la práctica, la implementación de algunos de los contrastes presentados a lo largo de toda la sección 3 requiere ciertos detalles que no especificamos entonces para no hacer farragosa la exposición con cuestiones que no son centrales para el entendimiento de los tests. Desarrollaremos ahora éstas para que el usuario de los contrastes pudiera implementarlos de forma completa. El lector poco interesado en aplicar los tests en la práctica puede pasar por alto este apartado. No obstante, debemos advertir que la comprensión de algunos de estos asuntos es relevante para interpretar de forma correcta y completa los resultados empíricos sobre las propiedades de los tests.

Frecuencias muestrales nulas Como se ha ido mencionando al exponer los tests, Mult2 y RV-p no podrían ser implementados si, dada la muestra utilizada, algún elemento de ciertos vectores de frecuencias relativas muestrales, necesarios en el cálculo de los estadísticos de contraste, resultara nulo (nos estamos refiriendo al vector $\overline{p}$ en el caso del test Mult2 y $\widehat{p}$ en el caso de RV-p). Para evitar tener que renunciar a la aplicación del contraste, nuestra propuesta es muy sencilla: sustituir la verdadera frecuencia relativa muestral nula por un valor muy pequeño, llamémosle ξ,[23] reconstruyendo convenientemente el resto de elementos del vector de frecuencias relativas de modo que su suma siga siendo uno. En concreto, sea r el vector de dimensión l de frecuencias relativas ($r = \overline{p}$ en el test Mult2, y $r = \widehat{p}$ en RV-p) el mecanismo exacto que estamos empleando es el siguiente: sustituir cada elemento $r_i = 0$ por $r'_i = \xi$ y cada $r_j > 0$ por $r'_j = r_j - \frac{s_1 \xi}{l - s_1}$, siendo s_1 el número de elementos de r que resultaron nulos.[24]

Creemos que éste es el modo de lograr aplicar el test distorsionando lo menos posible la información de las observaciones muestrales, sin tener que reducir la dimensión del vector p y, en consecuencia, sin tener que reformular la partición del dominio de datos y previsiones, la función de pérdida g y la función f de comparación entre las pérdidas. Según la forma en que se han elegido estas funciones, f puede generar K posibles valores distintos, pero, si p_i fuera efectivamente 0, significaría que las definiciones realizadas son incorrectas y deben rehacerse, ya que f solo debería producir $K - 1$ valores distintos. Para chequear el buen comportamiento del método, hemos comparado los resultados de estos dos tests para los ejercicios de simulación de la próxima sección usando dos estrategias alternativas: por un lado, empleando el método que proponemos aquí y, por otro, aplicando cada contraste solo si la muestra no generó este problema (es decir, solo se contabiliza el resultado de Mult2 (RV-p) en muestras para las que $\overline{p}_i > 0 \ \forall i$ ($\widehat{p}_i > 0 \ \forall i$)).

Pues bien, en la Tabla 15 del Apéndice B se adjuntan los resultados de Mult2 para la doble estrategia mencionada. Puede comprobarse que el tamaño y la potencia del test no se ven prácticamente modificados por emplear el procedimiento sugerido en la mayoría de casos. Solo cuando la correlación entre los dos errores de previsión es muy elevada y la muestra muy corta ($\rho = 0{,}9$ y $T = 8, 16$) se observa un empeoramiento del tamaño del test y, especialmente, una fuerte pérdida de potencia.[25] Por su parte, la Tabla 16 del Apéndice B presenta la misma comparativa pero para RV-p. En este caso, las propiedades del test

[22] Por supuesto, un análisis de las probabilidades estimadas $\widehat{p}$ sí permitiría obtener tal conclusión.

[23] En base a los resultados de simulaciones realizadas, recomendamos usar un valor de ξ que esté en el intervalo $(0, 0{,}001]$.

[24] En el caso del test RV-p, para hacer coherente la modificación en las frecuencias relativas y en las absolutas, añadimos un segundo paso: sustituir todas las frecuencias absolutas muestrales n_s por $n'_s = T\widehat{p}_s$ (para $s = 1, ..., K$). Una vez hecho esto, se aplica el contraste con los nuevos valores $\widehat{p}'_s$ y n'_s.

[25] En el caso $T = 8$, $\rho = 0{,}9$ el tamaño cae 2 puntos, y la potencia empeora 25. En este escenario, la probabilidad de que se verifique la condición $n_i + n_{K-i+1} = 0$ (que son los casos en que entra en juego nuestro método) es muy elevada (ver Tablas 11 y 13). La cifra de potencia alta (la que corresponde al caso simbolizado en la Tabla 15 por $\xi = 0$) se obtiene contabilizando los rechazos solo sobre los casos en que la citada condición no se verifica. Esto descarta, por ejemplo, los muchos casos en los que $\overline{p}_q = 1$ y $\overline{p}_i = 0 \ \forall i \neq q$ (es decir, todas las diferencias de pérdidas son cero), situaciones en las que debería no rechazarse la hipótesis nula. Por tanto, muy posiblemente, la potencia estimada en los casos $\xi = 0$ está sesgada al alza. Habría que distinguir qué parte del empeoramiento de los resultados es achacable al método, y qué parte a la distorsión de la cifra de SAP producida por la "selección" de casos a computar, que no es aleatoria.

mejoran considerablemente por el uso del procedimiento que sugerimos, ya que contabilizar el resultado del contraste solamente en situaciones donde todas las frecuencias son positivas lleva a infraestimar significativamente la probabilidad de RV-p de rechazar de la hipótesis nula.

Este mecanismo de corrección en los casos de frecuencias nulas logra, sin perjuicio en las propiedades de tamaño y potencia de los tests, que ambos sean siempre computables. La probabilidad de que, en caso de no aplicar ningún procedimiento para solucionarlo, estos contrastes no pudieran llevarse a cabo debido a este problema es muy elevada. Para adquirir conciencia de la relevancia de este asunto, hemos estimado dicha probabilidad en los experimentos de Monte Carlo de la sección a continuación, simplemente implementando los tests sin usar método alguno de corrección y contabilizando los casos en que no pudieron llevarse a cabo. Las estimaciones se muestran en las Tablas 11-14 del Apéndice B. De esas cifras se desprende que el problema era muy grave para el test Mult2 cuando las muestras son muy cortas, y dramático en el caso de RV-p, hasta el punto de que, si no se trataba este asunto de las frecuencias nulas, dicho test hubiera tenido que descartarse como procedimiento razonable para contrastar comparación de capacidad predictiva en muestras de longitud inferior a 50 observaciones.[26]

También el test Mult2-aprx puede tener un problema relacionado con frecuencias nulas. Tal y como se comentó en 3.2.3, dicho contraste no podría aplicarse si *todas* las frecuencias relativas son nulas menos la central (es decir, si $\widehat{p}_s = 1$ y $\widehat{p}_i = 0, i \neq s$), puesto que, entonces, el denominador del estadístico de contraste se anula. Por analogía con las situaciones de frecuencias nulas de Mult2 y RV-p, podría solucionarse el caso utilizando el mismo procedimiento. Aplicando dicho método para el cálculo de las frecuencias relativas $\widehat{p}$, se obtiene inmediatamente que el denominador del estadístico de contraste $M2_{apx}$ ya no se anula, y que $M2_{apx} = 0$, de modo que el test no rechazará la hipótesis nula, precisamente como hubiera sido deseable. Fijémonos que, ahora, la aplicación del procedimiento se limita al caso en que $\widehat{p}_i = 0 \ \forall i \neq s$, mientras en Mult2 y RV-p, bastaba un solo i para el que $\widehat{p}_i + \widehat{p}_{K-i+1} = 0$ y $\widehat{p}_i = 0$, respectivamente, para que necesariamente hubiera que aplicar el método. El hecho de que la aplicación del método sea mucho menos frecuente en Mult2-aprx que en los tests anteriores puede comprobarse en las Tablas 11-14.

Regiones críticas de tamaño α en contrastes discretos. La aleatorización Otro asunto relevante en la práctica es la *imposibilidad de construir regiones críticas de tamaño* α y la consiguiente posible solución, la *aleatorización*. Este problema afecta solamente a tests cuyo estadístico de contraste es una variable aleatoria discreta. En nuestro caso, por tanto, concierne a los tests de Signos, Wilcoxon y Mult2.

La descripción del problema y de la solución teórica (construcción de tests aleatorizados) son bien conocidas en estadística, y se presentaron en el Apéndice A del Capítulo 1 de la Tesis. En nuestras simulaciones, los tres tests citados se implementan usando el procedimiento de aleatorización, tal y como es recomendable para este tipo de experimentos. De no emplearse dicho método, los resultados de las simulaciones mostrarían sistemáticamente sesgo en tamaño en los tests discretos, conclusión errónea al ser provocada simplemente por haber ignorado la naturaleza discreta del estadístico de contraste y haber forzado a que el contraste tome siempre la misma decisión (rechazo o no rechazo) en los casos "conflictivos" (reléase el apartado sobre aleatorización del Apéndice A del Capítulo 1 de esta Tesis). En cambio, en la aplicación práctica, es decir, cuando se usa el test en muestras reales en vez de en ejercicios de simulación, los contrastes de distribución discreta no suelen ser implementados bajo su versión aleatorizada, sino que, en los casos de ambigüedad en la decisión del contraste −que son los que resuelve la aleatorización−, suele ser preferible que el usuario tome él mismo la decisión, redefiniendo el riesgo del error de tipo I que desea asumir para el contraste. En el Apéndice A del Capítulo 1 se detallaban también estas cuestiones.

En los resultados del capítulo actual, la aleatorización juega un papel mucho más relevante que el que tuvo en el Capítulo 1. En este epígrafe nos limitamos a presentar de forma más o menos breve cuatro cuestiones relativas al procedimiento de aleatorización en el contexto que nos ocupa, y será en el apartado E.1 del Apéndice E, donde éstas se expliquen con más detalle. Las tres primeras pueden incluso ser pasadas por alto por lectores no interesados en la implementación concreta de los tests. Sin embargo, queremos enfatizar la necesidad de una lectura atenta del cuarto comentario y de la ampliación de éste en el Apéndice E, puesto que es importante para la interpretación correcta de los resultados que los tests Signos, Wilcoxon y Mult2 obtengan en las simulaciones:

[26] Recuérdese que el problema de frecuencias nulas aparece en Mult2 cuando existe i tal que $n_i + n_{K-i+1} = 0$, mientras para RV-p basta que $n_i = 0$, condición ésta segunda bastante menos exigente que la primera. Ésta es la razón de que dicho problema sea aún de mucho mayor gravedad en el segundo test que en el primero.

a) Para mostrar las propiedades estadísticas de los tests discretos, en las simulaciones utilizamos sus versiones aleatorizadas, como ya se había dicho. No obstante, existe una alternativa, que emplearon Diebold y Mariano (1995) para los tests de Signos y Wilcoxon, consistente en fijar como niveles de significación en los experimentos de simulación valores de probabilidad que aparezcan en las tablas de la distribución del contraste, es decir, probabilidades correspondientes a puntos de discontinuidad de la función de distribución. Por ejemplo, para el test de Signos, Diebold y Mariano (1995) fijan como nivel de significación los valores 0,25, 0,1408 y 0,1536, para longitudes muestrales 8, 16, y 32, respectivamente, en vez del habitual nivel $\alpha = 0{,}10$, que emplean para los contrastes de distribución continua. Si se obra de esta manera, obviamente, la aleatorización ya no es necesaria, pese a que los estadísticos de los tests sean variables aleatorias discretas. Sin embargo, por argumentos tanto conceptuales como prácticos, no consideramos que sea una forma de proceder razonable en nuestro caso. Para conocer dichos argumentos, consúltese el apartado E.1 del Apéndice E.

b) En principio, está garantizado que el tamaño empírico de un test discreto aleatorizado será igual al teórico. Dicha propiedad se cumple por la propia construcción del método: se fuerza a que el parámetro del experimento de Bernouilli empleado para aleatorizar tome el valor adecuado para que el tamaño del test sea el teórico.[27] Sin embargo, se está suponiendo que la distribución de contraste es conocida con exactitud. En nuestro caso, ésta es la situación solamente del test de Signos, pero no de la distribución de Wilcoxon ni de la de Mult2. En el caso de Mult2, porque la distribución del estadístico de contraste solo se podría caracterizar completamente si el vector paramétrico p fuera conocido, en vez de estimarse, tal y como ya se expuso en la presentación del test. En el caso de Wilcoxon, los motivos se ofrecen en el apartado E.1 del Apéndice E. Entonces, en estos dos contrastes, el cálculo del parámetro de la variable de Bernouilli se está haciendo bajo una distribución que no es exactamente la verdadera, y de ahí que puedan producirse discrepancias entre el tamaño empírico y el fijado. En Wilcoxon, la distribución empleada en la implementación del test debe ser muy próxima a la verdadera, por lo que el sesgo en tamaño también debe ser pequeño. Sin embargo, en Mult2, cuando la muestra es muy corta y el valor numérico del parámetro p_s elevado, la distribución estimada puede diferir notablemente de la verdadera, tal y como se ha afirmado en el tercer apartado de 3.2.2, y la cuantía del sesgo podría ser considerable, pese a la aleatorización.

c) El problema de la región crítica se atenúa conforme mayor es el número de elementos del soporte del estadístico de contraste, ya que su distribución se aproxima entonces a una continua. En el caso de Signos y Wilcoxon, la cardinalidad del soporte depende únicamente de la longitud muestral, mientras en Mult2 también depende del número de pérdidas definido, que se denotó por K. En general, puede afirmarse que si se manejan muestras relativamente grandes, el problema en cuestión tendrá muy poca relevancia en cualquiera de los tres tests, por lo que, en dicho caso, el usuario puede ignorarlo por completo en una aplicación práctica. En el apartado E.1 del Apéndice E se muestra algún detalle más sobre este asunto.

d) Como se dijo arriba, en la práctica, los tests discretos no se aleatorizan, sino que su implementación consiste en tomar la decisión pertinente (rechazo o no rechazo) en aquellos casos en que no haya conflicto, y simplemente informar al usuario de los casos conflictivos y de la redefinición del riesgo de error de tipo I que tendría que asumir para tomar la decisión en dichos casos. Utilizando la notación del apartado sobre aleatorización en el Apéndice A del Capítulo 1, en un caso conflictivo (muestras en las que el valor del estadístico de contraste resultó $\lambda = \lambda_d$), el usuario deberá elegir entre no rechazar la hipótesis nula a nivel de significación $F_{(b)}$ o rechazarla a nivel de significación $F_{(d)}$. Entonces, parece claro que hay dos propiedades que son deseables en un test discreto, en la práctica:

– Por un lado, que el número de casos "conflictivos" sea lo menor posible, es decir, que la probabilidad del suceso $\{\lambda = \lambda_d\}$ (que es la probabilidad de tener que usar aleatorización, dada una muestra) sea pequeña.

– Por otro, si la distancia entre los niveles de significación $F_{(b)}$ y $F_{(d)}$ es pequeña, el test puede considerarse informativo, pero si es grande, la información aportada por el contraste no es muy valiosa. Por ejemplo, supóngase que efectivamente ocurre $\lambda = \lambda_d$ y resulta que $F_{(b)} = 0{,}05$ y $F_{(d)} = 0{,}35$, el test solo estaría proponiendo elegir entre no rechazar la hipótesis nula al 5 % o rechazarla al 35 %, manteniendo un nivel de ambigüedad elevado. Dicha distancia es precisamente $F_{(d)} - F_{(b)} = P(\lambda = \lambda_d)$, es decir, es igual, de nuevo, a la probabilidad de que ocurra el suceso $\{\lambda = \lambda_d\}$.

[27] Repásese Apéndice 1 del capítulo 1, en la parte que describe el mecanismo de aleatorización de tests de naturaleza discreta.

En resumen, además del tamaño y potencia, hay una propiedad fundamental que valorar en los tests discretos: que la probabilidad $P(\lambda = \lambda_d)$ asociada sea pequeña. Denotaremos dicha probabilidad por P_d. Como se dijo en el punto b), la aleatorización garantiza insesgadez en tamaño para cualquier test discreto cuya distribución de contraste sea conocida con exactitud. Sin embargo, si este resultado se ha obtenido gracias a aleatorizar frecuentemente, es decir, con un elevado P_d, la propiedad pierde valor, al ser en realidad solo teórica. Y, por supuesto, entre dos contrastes discretos con resultados de tamaño y potencia similares, será preferible el que presente menor P_d.

Por todo lo argumentado en este punto, cuando llevemos a cabo nuestros experimentos de simulación para estimar el tamaño de los contrastes, también mediremos la probabilidad P_d en los tres tests discretos. Por razones teóricas, sospechamos que los resultados para Mult2 respecto a P_d van a ser mejores que los de Signos y Wilcoxon. Estas razones y otras cuestiones relacionadas con la estimación de P_d se comentan en el apartado E.1 del Apéndice E.

Diferenciales iguales en Signos, Wilcoxon y DM En teoría, los contrastes de Signos y Wilcoxon no pueden llevarse a cabo cuando todos los diferenciales d_t de la muestra son nulos. La razón es que estos tests proceden "limpiando" la muestra de aquellos valores $d_t = 0$, para luego calcular de sus estadísticos de contraste, de modo que se quedarían sin muestra si $d_t = 0\ \forall t$. Por su parte, el test DM tampoco podrá llevarse a cabo si todos los diferenciales toman el mismo valor (sea cero o no), suponiendo que la estimación de la varianza de su estadístico de contraste es la habitual y que nos encontramos en el caso $h = 1$. Esto ocurre porque dicha varianza sería nula si $d_t = \overline{d}\ \forall t$, siempre que se estime por $\frac{1}{T}\sum_{t=1}^{T}(d_t - \overline{d})^2$. Este resultado ya se había establecido en la propiedad [Prop 7] del apartado 3.2.3.

Cuando d_t es una variable continua, obviamente, esta situación es muy poco verosímil, y el asunto no tiene ninguna relevancia, pero en un contexto discreto, la probabilidad de que suceda no es nula e incluso es elevada en casos de alta correlación contemporánea entre los dos conjuntos de previsiones. En las Tablas 11-14 del Apéndice B se presentan estimaciones de dicha probabilidad, para los escenarios de simulación utilizados en los análisis de Monte Carlo de la sección próxima.

Precisamente, la decisión deseable de los contrastes en este tipo de situaciones es obvia: no rechazar la hipótesis nula cuando los diferenciales son todos nulos y rechazarla si son todos iguales pero no nulos. Por tanto, nuestra implementación de los tests de Signos y Wilcoxon en el contexto discreto y del test DM en el caso $h = 1$ será:

– Si $d_t = 0\ \forall t$, no rechazar la hipótesis nula, directamente.

– Si $d_t = \overline{d} \neq 0\ \forall t$, DM rechazará directamente (Signos y Wilcoxon no necesitaban ninguna corrección en esta situación).

– En cualquier otro caso, aplicar los contrastes normalmente, es decir, tal y como se han explicado en la sección 3.1.

Es fundamental tener en mente que la implementación que acabamos de especificar, si bien garantiza la computabilidad de los tests de Signos y Wilcoxon para cualquier muestra, no lo hace para el test DM. Asegura que el contraste podrá ejecutarse siempre siempre que $h = 1$. En cambio, para $h > 1$, existe la posibilidad de que la estimación de la varianza del estadístico de contraste sea negativa, en cuyo caso el test DM no podría aplicarse. El motivo es que la varianza incorporará autocovarianzas muestrales de d_t de órdenes entre $-(h-1)$ y $h-1$. En los ejercicios de simulación se obtendrán resultados sobre la frecuencia con la que ocurren dichas situaciones (Cuadro 2, en sección 4).

Autocorrelación en la muestra. Cotas de Bonferroni A excepción de DM, todos los contrastes presentados a lo largo de la sección 3 están derivados bajo el supuesto de que la serie $\{zz_t\}_{t=1}^{T}$ no presenta autocorrelación, pero es frecuente que esta condición no se verifique en la práctica.[28] Antes de seguir adelante exponiendo el mecanismo de corrección a este problema, debe notarse que, aunque, obviamente, la serie $\{zz_t\}_{t=1}^{T}$ hereda autocorrelación cuando las muestras de datos y previsiones presentan, a su vez, autocorrelación, el grado de ésta puede atenuarse mucho en un ámbito de pérdidas discretas, como el nuestro. En el Cuadro 4 de la sección siguiente podrá constatarse este hecho.

[28] Recuérdese que en los tests Signos y Wilcoxon, las variables zz_t son los diferenciales d_t, mientras en los contrastes Mult2, Mult2-aprx y RV-p esto no tiene por qué ser así necesariamente, pero es la implementación más habitual.

Cuando se viola el supuesto de no autocorrelación, el procedimiento adecuado consiste en usar contrastes múltiples con nivel de significación definido por la cota de Bonferroni en el sentido sugerido en otro contexto por Campbell y Ghysels (1995), tal y como hacen Diebold y Mariano (1995) para Signos y Wilcoxon. El procedimiento es el que sigue:

Sea r el orden de la autocorrelación de la muestra $\{zz_t\}_{t=1}^{T}$, se trata de separar la muestra en $r+1$ submuestras no autocorreladas, de la forma $\left\{zz_1, zz_{(r+1)+1}, zz_{2(r+1)+1}...\right\}$, $\left\{zz_2, zz_{(r+1)+2}, zz_{2(r+1)+2}...\right\}$, ..., $\left\{zz_{r+1}, zz_{(r+1)+r+1}, zz_{2(r+1)+r+1}...\right\}$. Cualquier test que requiere no autocorrelación puede implementarse para un nivel de significación α sobre una muestra autocorrelada aplicando el test sobre cada una de las $r+1$ submuestras usando nivel de significación $\alpha' = \alpha/(r+1)$ y rechazando la hipótesis nula si ésta fue rechazada en una o más de las submuestras. Bonferroni demuestra que el tamaño del test conjunto será menor o igual que α. Usando este procedimiento, los contrastes serán algo más costosos de implementar y requerirán una identificación previa del orden de autocorrelación de la secuencia de diferenciales de pérdida. No obstante, también el test DM necesita dicha identificación para su aplicación, ya que se requiere para estimar $f_d(0)$.

El de Bonferroni será el método que utilizaremos en nuestra implementación de los contrastes expuestos a lo largo de toda la sección 3 (salvo en el test DM), en todos aquellos ejercicios de simulación de la sección siguiente que introduzcan autocorrelación en la muestra.

4. Análisis de Monte Carlo

No conocemos las propiedades estadísticas en muestras finitas de los contrastes que hemos presentado ni tampoco las de los tests DM, Signos y Wilcoxon cuando se usa una función de pérdidas g discreta, del tipo a las definidas en el apartado 2.2. Por ello, realizamos experimentos de Monte Carlo para determinar el tamaño y potencia de estos seis tests, bajo distintas condiciones sobre las pérdidas $z_{t,i}$ y sobre el proceso $zz_t = f(z_{t,1}, z_{t,2})$. En la práctica, los analistas se encuentran con muestras muy reducidas de previsiones, por lo que hemos centrado el estudio en longitudes muestrales muy cortas, $T = 8, 16, 24, 32, 40, 48$.

Aunque se ha destacado a lo largo del trabajo el interés práctico de poder definir la función g en términos de los dos inputs $y_t, v_{t,i}$ en vez de simplificar la información de ese par en el error de previsión $e_{t,i}$, consideramos que lo apropiado a la hora de llevar a cabo el estudio de Monte Carlo para obtener estimaciones del tamaño y potencia de los tests será utilizar una función de pérdidas $g(e_{t,i})$, tal y como hicieron en sus ejercicios Diebold y Mariano (1995), Harvey et al (1997) y Dell'Aquila y Ronchetti (2004), por ejemplo. Pensamos que ésta es la implementación adecuada para los experimentos tanto por su sencillez como, sobre todo, por la flexibilidad que permite para introducir escenarios diferentes en términos de autocorrelación en zz_t y de correlación contemporánea entre las pérdidas $z_{t,i} = g(y_t, v_{t,i})$.

4.1. Diseño del Experimento

1) Los experimentos que definimos para evaluar las propiedades de los contrastes seguirán un diseño similar al que emplearon Diebold y Mariano (1995) para estimar el tamaño del test DM. Se generan realizaciones $\{e_{t,1}, e_{t,2}\}_{t=1}^{T}$ de un proceso gaussiano bivariante de errores de previsión, con distintos grados de autocorrelación y correlación contemporánea, que se define por

$$e_{t,i} = (1+\theta^2)^{-1/2}\left(v_{t,i} - \theta v_{t-1,i}\right),\ v_{t,i} \overset{iid}{\sim} N(0, \sigma_i^2),\ i = 1, 2,$$
$$E(v_{t,1}v_{t,2}) = \rho\sigma_1\sigma_2;\quad E(v_{t,1}v_{s,2}) = 0,\ \forall t \neq s.$$

De este modo, los procesos $e_{t,i}$ verifican las siguientes propiedades:

$$(e_{t,1}, e_{t,2})' \sim N(0_{2\times 1}, \Sigma),\ \Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho \\ \sigma_1\sigma_2\rho & \sigma_2^2 \end{pmatrix},$$
$$e_{t,i} \sim N(0, \sigma_i^2),\ i = 1, 2,$$
$$\varphi_1(e_{t,i}) = -\frac{\theta}{1+\theta^2},\ i = 1, 2,$$
$$\eta(e_{t,1}, e_{t,2}) = \rho,$$

siendo $\varphi_1(e_{t,i})$ el coeficiente de autocorrelación de orden 1 para $e_{t,i}$ y $\eta(e_{t,1}, e_{t,2})$ el coeficiente de correlación contemporánea entre $e_{t,1}$ y $e_{t,2}$.

El procedimiento para generar el proceso $e_{t,i}$ es el mismo que en Diebold y Mariano (1995): Se generan primero T realizaciones del proceso $(u_{t,1}, u_{t,2})' \sim N(0_{2\times 1}, I_{2\times 2})$ y se premultiplican por el factor de Cholesky

$$\Gamma = \begin{pmatrix} \sigma_1 & 0 \\ \rho\sigma_2 & \sigma_2\sqrt{1-\rho^2} \end{pmatrix}.$$

Los nuevos errores transformados son $(v_{t,1}, v_{t,2})' \sim N(0_{2\times 1}, \Sigma = \Gamma\Gamma')$ y presentan las varianzas y correlación contemporánea deseadas. El segundo paso consiste en añadir la autocorrelación de primer orden, definiendo $e_{t,i}$ como un proceso MA(1) por $e_{t,i} = (1+\theta^2)^{-1/2}(v_{t,i} - \theta v_{t-1,i})$. El factor $(1+\theta^2)^{-1/2}$ permite conservar las varianzas incondicionales deseadas. En nuestros experimentos, impondremos siempre varianza unitaria para $e_{t,2}$, es decir, se impone $\sigma_2^2 = 1$.

2) La función discreta $g(e_{t,i})$ para generar pérdidas $z_{t,i}$ que será empleada para los seis contrastes que evaluaremos consistirá en clasificar los errores de previsión en G+ ($e > +1$), G- ($e < -1$) y P ($e \in [-1, +1]$) y penalizarlos según (G2), función presentada en la sección 1. Elegimos ésta por ser una función sencilla y razonable para valorar errores de previsión. Los errores "grandes" se valoran asimétricamente según signo (más graves los negativos). Como se verá en la próxima sección, la influencia de estas elecciones (partición y asignación de pérdidas) sobre las conclusiones de los experimentos es pequeña, y en ningún caso juega papel alguno en la comparativa de propiedades entre los distintos tests.

Por su parte, para aplicar los tests RV-p, Mult2 y Mult2-aprx utilizaremos la función de comparación $f(z_{t,1}, z_{t,2})$ especificada en (F2), que coincide con la función $d_t = f(z_{t,1}, z_{t,2}) = z_{t,2} - z_{t,1}$ del test DM. De este modo, la implementación de los contrastes que presentamos y del test DM se realiza bajo las mismas funciones f y g, permitiendo comparaciones homogéneas.

3) Nuestros experimentos tratarán de estimar propiedades estadísticas de los tests bajo distintos niveles de autocorrelación en zz_t y correlación cruzada entre *las pérdidas* $z_{t,i}$. Nuestro mecanismo para generar estos escenarios es transmitir autocorrelación y correlación contemporánea a través de $e_{t,i}$, igual que se hace en Diebold y Mariano (1995). Sin embargo, debe quedar claro que, una vez filtrado $e_{t,i}$ por la función g en (G2), las variables $z_{t,i}$ y zz_t están heredando estas características solo parcialmente, es decir, la autocorrelación de orden 1 en zz_t y la correlación contemporánea entre $z_{t,1}$ y $z_{t,2}$ no serán $\varphi_1(\theta)$ y ρ, respectivamente, sino bastante menores. En los Cuadros 4 y 5 se exponen las estimaciones que resultaron del propio experimento para estos parámetros.

4) Llevamos a cabo dos experimentos, uno para estimar tamaño y otro para estimar potencia. El diseño es exactamente el mismo en ambos casos, excepto en dos cuestiones:

− Varianza de errores de previsión: en el experimento para tamaño, se elige $\sigma_1^2 = 1$ y, por el contrario, en el de potencia, se elige $\sigma_1^2 = 3$, mientras en ambos casos imponemos $\sigma_2^2 = 1$. Por lo tanto, los conjuntos de previsiones 1 y 2 son de igual precisión en el experimento de tamaño, mientras el conjunto 2 es preferible al 1 en el de potencia.

− En consecuencia, las hipótesis alternativas que definimos para los contrastes son distintas:

a) En tamaño: "los conjuntos tienen *distinta* capacidad predictiva". Esto se traduce en $H_1 = \neg H_0^{(SW)}$ (Signos y Wilcoxon), $H_1^{(1d)}$ (DM, Mult2 y Mult2-aprx) y $H_1^{(2)}$ (RV-p). La región crítica de los tests es la unión de las colas inferior y superior de las distribuciones de contraste, salvo en RV-p.

b) En potencia: "el conjunto 2 tiene *mayor* capacidad predictiva que el 1". Esto se traduce en $H_1 = med(d_t) < 0$ (Signos y Wilcoxon) y $bp < 0$ (DM, Mult2 y Mult2-aprx). Ahora, la región crítica es solo la cola inferior de las distribuciones. RV-p no admite una hipótesis alternativa distinta a $H_1^{(2)}$, por lo que la implementación del test es la misma en el experimento para potencia que en el de tamaño.

En vez de estimar la potencia a través de la probabilidad de rechazo bajo hipótesis nula falsa, emplearemos el concepto de "potencia ajustada a tamaño" (SAP), como es habitual en la literatura estadística. Como es bien sabido, dicha medida evita la distorsión que presenta la comparación entre la potencia de un grupo de contrastes cuando alguno de ellos es sesgado en tamaño. Un test que rechaza su hipótesis nula pocas veces cuando es cierta (sesgo por defecto en tamaño), tiende a hacer lo mismo cuando es falsa (infraestimando potencia), y viceversa, lo que deriva en una imagen irreal favorable en términos de potencia a los tests sesgados por exceso en tamaño. En el Apéndice A se define el concepto de SAP y algunas cuestiones para la implementación del mismo en los tests que nos ocupan.

Debemos hacer notar que, teóricamente, existe un problema para la comparación de las potencias de los contrastes evaluados, debido a que las hipótesis nulas no son exactamente iguales. Es decir, bajo nuestro escenario $\sigma_1^2 = 3$ y $\sigma_2^2 = 1$, ¿el grado de falsedad de las hipótesis $H_0^{(1)} = bp = 0$ (DM, Mult2, Mult2-aprx), $H_0^{(SW)} = med(d_t) = 0$ (Signos y Wilcoxon) y $H_0^{(2)} = p_i = p_{K-i+1}$ (RV-p) es el mismo en los tres casos?. En general, no, obviamente. En el diseño de nuestros experimentos, la distribución de d_t es simétrica respecto a 0 y, por tanto, la media y la mediana coinciden en el punto $d = 0$. En consecuencia, $med(d_t) = E(d_t) = bp = 0$, por lo que las dos primeras hipótesis mencionadas resultarán exactamente la misma, y las potencias de todos los tests son comparables entre sí, salvo RV-p. En cualquier caso, todos los tests pretenden contrastar si los conjuntos de previsiones tienen la misma capacidad predictiva y ésta es su hipótesis implícita, por lo que la comparación numérica de las potencias de los tests, incluido RV-p, puede considerarse correcta, en ese sentido.

5) El experimento de tamaño y el de potencia se desdoblarán, a su vez, en dos ejercicios:

– El primero fija la condición $\theta = 0$, es decir, supone horizonte de previsión $h = 1$.[29] Para este caso, se especifican tres escenarios de correlación contemporánea, $\rho = 0$, 0,5, 0,9.

– En el segundo, se impone $\theta = -0{,}5$, suponiendo $h = 2$. Ahora se emplean dos escenarios de correlación contemporánea, $\rho = 0$, 0,5. Se ha comprobado que los resultados no varían si $\theta = -0{,}9$ (téngase en cuenta que la diferencia entre $\varphi_1(\theta = -0{,}5) = 0{,}4$ y $\varphi_1(\theta = -0{,}9) \simeq 0{,}5$ es pequeña). En el diseño $\theta = -0{,}5$, no se incluye el caso $\rho = 0{,}9$, porque en situaciones de correlación cruzada tan alta y pérdidas de tipo discreto, la autocorrelación de zz_t disminuye en tal cuantía que la decisión de aplicar o no el método de contrastes múltiples con cota de Bonferroni se vuelve confusa (véase Cuadro 4).

Este diseño partido en dos según θ (análogamente, según h) coincide con el empleado por Harvey et al (1997) y Dell'Aquila y Ronchetti (2004) y, por el contrario, difiere del que presentaron Diebold y Mariano (1995). Estos últimos no identificaban $\theta = 0$ con $h = 1$ y $|\theta| > 0$ con $h > 1$, sino que su diseño incluía $\theta = 0$ y $|\theta| > 0$ en el mismo experimento, pero todos los casos seguían correspondiéndose con $h = 2$. De este modo, su implementación del test DM siempre era la misma, usando $L(T) = 1$ (es decir, incluyendo siempre la autocovarianza muestral de orden uno $\widehat{\gamma}_d(1)$ en el cálculo de la varianza de $\overline{d}$ (véase apartado 3.1.1)), independientemente de que existiera realmente autocorrelación o no, lo que no parece razonable. Nosotros, en cambio, calculamos S_1 usando $L(T) = 0$ si $h = 1$ y $L(T) = 1$ si $h = 2$. La aparente robustez del test DM a la presencia de autocorrelación mostrada en Diebold y Mariano (1995) está relacionada con este asunto. De utilizarse nuestro diseño, *los resultados de DM en muestras cortas sí dependen de la autocorrelación* o, equivalentemente, de h. Nosotros mostraremos este resultado para función de pérdida g discreta, resultado que ya había sido mostrado por Harvey et al (1997) y Dell'Aquila y Ronchetti (2004) para la función de pérdida cuadrática $g(e_{t,i}) = e_{t,i}^2$.

6) En los ejercicios donde $|\theta| > 0$, los tests de Signos, Wilcoxon, Mult2, Mult2-aprx y RV-p se aplican utilizando el método de Bonferroni descrito en el cuarto apartado de 3.2.5, para evitar los problemas derivados de la existencia de autocorrelación en la variable zz_t. Haremos una excepción: si $T \leq 16$, dicho método no se empleará en ninguno de los cinco tests anteriores, aunque exista autocorrelación, implementándose los tests como si no la hubiera. Se ha comprobado que la probabilidad de rechazar la hipótesis nula que presentan estos tests es muy inferior a la teórica en las submuestras de tamaño $T/2$ que resultan (4 y 8 datos, respectivamente), que serían las utilizadas en caso de aplicación del método de Bonferroni. Recuérdese que este método solo garantiza que el tamaño del test conjunto no supera el nivel de significación teórico α, pero puede ser menor. Nuestras pruebas sugieren que, en este contexto, el tamaño del test conjunto es claramente inferior a α en muestras inferiores a 8 datos, de modo que el procedimiento no funciona bien en tales casos. Por ello, también recomendamos al usuario de estos contrastes que siga esta misma directriz en una posible implementación práctica: no emplear Bonferroni si la longitud de la muestra no alcanza las 16 observaciones.

7) Salvo DM, todos los tests son computables para cualquier muestra. El test Mult2-aprx lo es por construcción, salvo en un caso concreto. En el primer apartado de 3.2.5 se explicaron mecanismos sencillos para que los contrastes Mult2, Mult2-aprx y RV-p tuvieran garantizada su aplicación siempre. Fijamos el valor del parámetro ξ que se empleaba en dichos procedimientos en $\xi = 0,0001$. Los tests de Signos y Wilcoxon no serían computables, en teoría, cuando todos los diferenciales d_t de la muestra resultan nulos, en cuyo caso se obtendría una muestra final de longitud $T' = 0$. La solución es simple y se expuso en el tercer apartado de 3.2.5: directamente, no rechazar la hipótesis nula $H_0^{(SW)}$ en dichas situaciones. A su vez, el test DM no se podría computar cuando todos los diferenciales son iguales, sean nulos o no, ya que se anularía el denominador del estadístico de contraste S_1. También la corrección se mostró en el tercer apartado de 3.2.5: nuevamente, no rechazamos directamente la hipótesis nula si los diferenciales eran todos nulos, y la rechazamos si eran no nulos.

[29] Dicha afirmación se realiza bajo el supuesto de que, en virtud de la descomposición de Wold para el proceso a predecir, los errores de previsión óptimos a horizonte h presentan autocorrelación de grado máximo $h - 1$.

Con las correcciones empleadas, solo hay un caso en que uno de los tests puede no ser computable. Se trata del test DM, cuando la serie d_t presenta autocorrelación, algo que, en nuestras simulaciones, se restringe al caso $h = 2$ ($\theta \neq 0$). En tales situaciones, la varianza que aparece en el denominador de S_1 se estima incorporando las autocovarianzas muestrales de d_t de órdenes 0 y 1, y nada impide que dicha varianza pueda resultar nula o negativa.[30] En el Cuadro 2 se adjunta el porcentaje de veces en las que sucedió esta situación durante las simulaciones y en las que, por tanto, DM no pudo computarse.

8) En el apartado A.1 del Apéndice A se hacen constar ciertos detalles respecto a la implementación de los contrastes. Entre ellos, cabe destacar que los tests de Signos y Wilcoxon se aplican en sus versiones de muestras finitas siempre que $T \leq 64$ en el caso del contraste de Signos, y $T \leq 20$ en el de Wilcoxon. Además, debemos resaltar que se aplica, tanto en los ejercicios de tamaño como en los de potencia, el procedimiento de "aleatorización" en el test Mult2 y en las implementaciones de muestras finitas de los tests de Signos y Wilcoxon.[31]

9) En todos los ejercicios a continuación se ha empleado $\alpha = 0{,}10$ como nivel de significación y se han llevado a cabo 10000 repeticiones[32] para cada uno de los seis tamaños muestrales elegidos ($T = 8, 16, 24, 32, 48$). A diferencia de Diebold y Mariano (1995), nosotros llevamos a cabo la ejecución de los tests de Signos y Wilcoxon con el mismo valor de α que para el resto de tests aunque T sea pequeño, ya que utilizamos el mecanismo de "aleatorización" para la implementación de los contrastes.

4.2. Resultados

En las Tablas 5 a 8 se presentan los resultados obtenidos de los experimentos de simulación expuestos en el apartado anterior para los seis tests introducidos en secciones previas. Los resultados se obtienen para longitudes muestrales comprendidas entre 8 y 48, y para distintos contextos de autocorrelación y de correlación contemporánea en los errores de previsión. Los seis tests se implementan usando una función g de pérdidas discreta − la especificada en el apartado anterior − y nuestros contrastes Mult2, Mult2-aprx y RV-p usarán la misma función de comparación de pérdidas que DM, $f(z_{t,1}, z_{t,2}) = z_{t,2} - z_{t,1}$. En tal caso, recuérdese que DM y Mult2-aprx solo diferirán en la estimación de la varianza de su estadístico de contraste $\overline{d} = b\widehat{p}$, siempre que el horizonte de previsión sea uno.

Los resultados pueden sintetizarse del siguiente modo:[33]

1. En general, las diferencias entre los tests en términos de propiedades de tamaño y SAP son pequeñas. Aparentemente, solo RV-p resulta claramente inferior al resto, por ser menos exacto en tamaño y, sobre todo, por su escasa potencia. Probablemente, dicho resultado esté condicionado por la especificación de su hipótesis nula, diferente a las demás, como se mencionó más arriba. Entre los otros contrastes, los más exactos en tamaño resultan Signos y Mult2. Por su parte, Wilcoxon infraestima ligeramente tamaño de forma casi sistemática, mientras DM presenta sesgos importantes en muestras cortas cuando $\theta \neq 0$, y, finalmente, Mult2-aprx obtiene tamaños solo levemente peores que los de Mult2. Además, Mult2 es el test más potente cuando el horizonte de previsión es uno, mientras DM lo es en horizonte dos, salvo en muestras muy cortas. En cualquier caso, la superioridad de ambos frente al resto se produce casi siempre por una ventaja numérica pequeña.

A continuación abordamos un análisis más pormenorizado de los resultados obtenidos, para orientar sobre qué tests son preferibles en la práctica.

[30] En cambio, si $h = 1$, solo se incluye la autocovarianza de de orden 0, por tanto, no negativa. En dicho caso, la varianza de S_1 sería exactamente igual a 0 solamente si todos los diferenciales se anulan, situación que ya hemos indicado cómo corregir.

[31] Aunque, en teoría, Wilcoxon se aplica en su versión asintótica cuando $T > 20$, en la práctica, se está empleando la versión exacta (con aleatorización) del test en muchas muestras cuya longitud verifica dicha condición. La razón es que, tras limpiar la muestra inicial de $T > 20$ datos de observaciones $d_t = 0$, la longitud final T' de la muestra puede ser inferior a 20. Esto ocurre a menudo, ya que la probabilidad $P(d_t = 0)$ es elevada en este contexto (véase Cuadro 3).

[32] En el caso del test DM, en realidad, realizamos 10000 intentos pero el número de repeticiones efectivas es menor que 10000, cuando $h = 2$, debido a la aparición de varianzas nulas o negativas.

[33] Debe aclararse, antes de nada, que existe una explicación para las aparentes incoherencias que aparecen en la Tabla 6 respecto a las estimaciones del tamaño de los tests (salvo DM), que, en contra de lo esperable, no resulta una función monótona en T. En la Tabla 6, el tamaño de los tests está aumentando conforme crece T a partir de $T = 24$ y, sin embargo, disminuye entre $T = 16$ y $T = 24$. La razón es que los tests se implementan de distinto modo en los casos de longitudes $T = 8$ y $T = 16$, salvo DM. La implementación para estas situaciones se especificó y justificó en el apartado 4.1.

2. En el punto d) del segundo apartado de 3.2.5 se presentó una propiedad fundamental para los tests de naturaleza discreta, según la cual las probabilidades P_d (definidas también en dicho punto) deberían ser razonablemente pequeñas, y se justificó la relevancia de la propiedad. Pues bien, en la Tabla 19, que se presenta en el Apéndice E, constan las estimaciones de P_d para los tres contrastes discretos que nos ocupan, Signos, Wilcoxon y Mult2.[34] Los resultados son rotundamente favorables a Mult2, en relación a los otros dos tests. Para Mult2, P_d se mantiene acotada entre 0.03 y 0.10, y casi en todos los casos se encuentra por debajo de 0.07. Para Signos y Wilcoxon, P_d se sitúa en los intervalos $[0{,}09, 0{,}53]$ y $[0{,}01, 0{,}53]$, respectivamente, según el diseño.[35] En una buena parte de situaciones correspondientes a muestras de longitud inferior a 24 datos (en concreto, bien cuando $h = 1$ y ρ es alto, o bien cuando $h = 2$), la estimación de la probabilidad P_d correspondiente a estos dos tests se situó entre 0,20 y 0,53. Además, para $T = 8$ y $\rho \neq 0$, $\widehat{P}_d$ nunca fue inferior a 0,33. Eso significa que, en dichos casos, Signos y Wilcoxon son de muy poca utilidad para el usuario (revísese el punto d) del apartado sobre aleatorización en 3.2.5), y sus aparentemente buenos resultados en tamaño −que se corresponden precisamente con los casos en que se manifestaron superiores a Mult2− solo son producto del efecto que genera la aleatorización, pero no responden a propiedades reales.

Desde nuestro punto de vista, las claras diferencias en términos de P_d que existen entre Mult2 y los tests de Signos y Wicoxon, favorables al primero, son suficientes para considerar a éste preferible en la práctica respecto a los otros dos, pese a que, aparentemente, sus propiedades teóricas de tamaño y potencia resultaran similares. La razón de estas diferencias radica en que el soporte del estadístico de contraste de Mult2 es más denso que los soportes de los otros dos tests, característica que, obviamente, hace que la distribución F_{ZZ_n} sea más parecida a una distribución continua que las otras dos, y, por tanto, que P_d tienda a tomar valores más moderados.

3. Además del que se acaba de exponer, existe otro motivo, ya mencionado en la presentación del test en el apartado 3.1.2, por el que el test de Signos es poco recomendable: su potencia puede ser mucho menor que la del resto de contrastes en muchas situaciones. En nuestros experimentos de potencia, la probabilidad de obtener diferenciales de pérdidas negativos es mayor que la de obtener diferenciales positivos, es decir, $p_1 + p_2 > p_4 + p_5$ (véase Cuadro 3), situación que detecta el test de Signos. Sin embargo, si hubiéramos diseñado un experimento de potencia, donde se verificara $p_1 + p_2 = p_4 + p_5$ pero, sin embargo, la probabilidad de diferenciales negativos grandes fuera mayor que la probabilidad de diferenciales positivos grandes (es decir, $p_1 > p_5$), y viceversa en los diferenciales pequeños ($p_2 < p_4$), Signos tendería a no detectarlo, y la SAP estimada sería muy inferior a la del resto de contrastes. Por ejemplo, en un caso tan claro de superioridad prediciva a favor del segundo conjunto de previsiones como el asociado a los parámetros $p_1 = 0{,}5$, $p_2 = 0$, $p_3 = 0$, $p_4 = 0{,}5$, $p_5 = 0$, Signos no rechazaría la hipótesis nula.

4. Por su parte, el test de Wilcoxon, además de incurrir en el problema expuesto en el punto 2, también presenta un ligero sesgo en tamaño, de forma casi sistemática, pese a haberse implementado en su versión aleatorizada. La explicación consiste en el argumento expuesto en el punto 4 del apartado 3.1.2. Por último, otra posible crítica hacia el contraste de Wilcoxon es que su validez teórica descansa en el supuesto de simetría de la distribución de d_t, tal y como se mencionó en el comentario 5 del apartado 3.1.2. Dicho supuesto se cumple en nuestros experimentos, por lo que los resultados del test no se han visto afectados por esta limitación, pero podría no verificarse en muchas aplicaciones en la práctica.

5. En general, DM y Mult2-aprx presentan buenos resultados. Las propiedades teóricas obtenidas en el apartado 3.2.3 adelantaban ya conclusiones sobre la comparación entre estos tests si el horizonte de previsión es uno: la SAP debe ser exactamente la misma en ambos, y las discrepancias en tamaño, en principio, serán pequeñas, e intuíamos que favorables a Mult2-aprx, por cuestiones relativas a la eficiencia del estimador de la varianza del estadístico de contraste. Quedaba por constatar el valor numérico de estas posibles diferencias de tamaño en $h = 1$, y por analizar todas las propiedades de los dos tests en el caso $h = 2$, para el que no disponíamos de conclusiones teóricas.

[34] La razón de que las estimaciones de P_d se muestren en un Apéndice, pese a su importancia, descansa solo en cuestiones de espacio y ordenación lógica.

[35] Cuando la longitud muestral supera las 24 observaciones, las estimaciones $\widehat{P}_d$ para Wilcoxon se moderan notablemente. Sin embargo, esto solo es producto de la implementación del test que hemos aplicado en nuestras simulaciones, ya que se emplea su versión asintótica para $T > 20$. Este asunto se detalla en el apartado E.2 del Apéndice 5.

Las simulaciones muestran que, en general, DM es ligeramente sesgado en tamaño, por exceso, salvo cuando la muestra es pequeña y el horizonte de previsión es dos, donde el sesgo resulta inaceptable. En $T = 8$, el tamaño empírico de DM casi alcanza el 30 %, cuando el nivel de significación era 10 %. Éste *es el problema más relevante que encontramos en el test DM*, unido al hecho de que su computabilidad no esté garantizada en horizontes superiores a uno, en el contexto de funciones de pérdida discreta. En el Cuadro 2 se presentan los porcentajes de veces en que DM no pudo computarse por resultar no positiva la estimación de la varianza $2\pi f_d(0)$.

Por su parte, Mult2-aprx es muy exacto en tamaño en la mayoría de los escenarios de predicción, salvo en aquellos en los que todos los tests presentan sesgo en tamaño, a saber: $h = 2$ y muestras cortas ($T \leq 24$), o bien $h = 1$, muestras cortas y correlación muy alta ($T \leq 16$, $\rho = 0{,}9$). Creemos que la mayor exactitud en tamaño de Mult2-aprx respecto a DM cuando $h = 1$ se debe al uso de un estimador más eficiente para la varianza de la distribución asintótica empleada. Respecto a la potencia de Mult2-aprx, ésta es exactamente igual que la de DM en horizonte uno, siempre que se mida por SAP, en virtud del corolario de [Prop 9] establecido en 3.2.3. Sin embargo, Mult2-aprx es menos potente que DM en horizonte dos (la diferencia se estabiliza, aproximadamente, en 7 %), salvo en muestras muy cortas, donde la ventaja incluso es favorable al primero.

Resumiendo, dados los resultados obtenidos y teniendo en cuenta que está garantizada la computabilidad de Mult2-aprx, dicho test puede constituir una *alternativa razonable a DM* (en contextos de pérdidas discretas): cuando $h = 1$, Mult2-aprx es algo más exacto en tamaño que DM e igual de potente, mientras en $h = 2$, es claramente preferible en tamaño si la muestra es muy corta ($T \leq 16$), y peor en potencia en muestras más largas. Por tanto, Mult2-aprx tendría ligeramente mejores propiedades que DM en casos de no autocorrelación y sería especialmente preferible en situaciones de autocorrelación si la muestra es muy corta.

Cuadro 2. % No Computabilidad DM
Caso $h = 2$ ($\theta = -0,5$)

	Tamaño		Potencia	
	$\rho = 0,0$	$\rho = 0,5$	$\rho = 0,0$	$\rho = 0,5$
$T = 8$	4,4	3,1	6,0	5,3
$T = 16$	1,0	1,0	1,1	1,3
$T = 24$	0,3	0,2	0,2	0,4

Dado que el problema más grave asociado al test DM es el sesgo en su tamaño cuando la muestra es muy corta y además existe autocorrelación en los diferenciales de pérdidas, podríamos pensar en sustituir DM por la versión corregida de Harvey, Leybourne y Newbold (1997), que denotábamos en el apartado 3.1.1 por HLN, en vez de emplear Mult2-aprx. Como es bien sabido, es precisamente en los casos de longitudes muestrales pequeñas y horizonte de previsión superior a uno, cuando HLN mejora los resultados de tamaño de DM, suavizando el sesgo. Una restricción a esta forma de proceder es que HLN obliga a estimar $f_d(0)$ usando "lag-window" rectangular y $L(T) = h-1$ (repásese 3.1.1), de modo que, por ejemplo, no generaría ninguna corrección respecto a DM si el horizonte de previsión es uno y, sin embargo, existe autocorrelación en la muestra.[36] En cualquier caso, hemos llevado a cabo el mismo ejercicio de tamaño correspondiente al caso $\theta = -0{,}5$ ($h = 2$) descrito arriba, pero solamente para muestras $T = 8, 16$ y para los tests DM, HLN y Mult2-aprx, con la intención de comparar si la reducción del sesgo en tamaño que logra HLN respecto a DM es equiparable a la obtenida por Mult2-aprx. Los resultados del experimento se presentan en la Tabla 18 del Apéndice C. Como puede observarse en dicha Tabla, HLN logra disminuir aproximadamente cinco puntos el tamaño de DM en casi todos los casos, pero sigue lejos del teórico. Esto es especialmente notable en aquellos casos en los que existe correlación contemporánea entre las dos series de errores de previsión y, además, la muestra es solo de 8 observaciones, situación frecuente en la práctica. Los resultados de Mult2-aprx siguen siendo mejores que los de HLN.

[36] Situación que podría ocurrir porque los errores de previsión no fueran óptimos o las características del proceso estocástico que se trata de predecir no permitieran aplicar la descomposición de Wold.

6. En general, Mult2 *es el contraste con mejores propiedades* de los evaluados, aunque sus ventajas frente al resto no son grandes. En ausencia de autocorrelación, resulta el test de mayor potencia (Tabla 7), mientras su tamaño se ajusta correctamente al teórico para todos los niveles de correlación cruzada y todas las longitudes muestrales, salvo en el caso $T \leq 16$, $\rho = 0{,}9$.

Cuando se introduce autocorrelación, Mult2 es ligeramente menos potente que DM cuando $T > 16$. Por el contrario, cuando $T \leq 16$, Mult2 es claramente más potente que DM y, lo que es más relevante, su tamaño (que deja de ser exacto) es mucho más próximo al teórico que el de aquel. Los tests de Signos y Wilcoxon aparentan mayor precisión en tamaño, pero sus buenos resultados se corresponden con propiedades más teóricas que reales, como se explicó en el punto 2 de este apartado.

En este contexto, las ventajas de Mult2 frente a DM en términos de precisión en su tamaño proceden del hecho de que *Mult2 está utilizando una estimación* (que, según revelan las simulaciones, es buena) *de la distribución exacta* del estadístico $\overline{d} = b\widehat{p}$, mientras DM emplea su *distribución asintótica*.

7. Interesa analizar los resultados de Mult2 y Mult2-aprx comparativamente, para comprobar si el segundo puede ser una buena aproximación del primero, resultado que sería muy útil cuando T es grande, situación en la que el coste computacional de Mult2 se convierte en excesivo. En términos de tamaño, Mult2-aprx prácticamente obtiene los mismos resultados que Mult2. Este hecho era esperado, puesto que, tal y como se mostró en la Figura 1 del apartado 3.2.3, la distribución Normal (usada por Mult2-aprx) constituye una muy buena aproximación de la distribución exacta (usada por Mult2) del estadístico de contraste $b\widehat{p}$ empleado por ambos tests, incluso en muestras muy cortas. Por el contrario, se observan diferencias en potencia entre ambos tests, favorables a Mult2.

8. El tamaño de los tests no es robusto a la presencia de autocorrelación en muestras cortas, aun empleando el método de Bonferroni, como puede verse al comparar las estimaciones de las Tablas 5 y 7. Al introducir autocorrelación ($h = 2$), aparecen sesgos por defecto en el tamaño de todos los tests menos DM, donde el sesgo es por exceso.[37] La no robustez del test DM a la autocorrelación puede resultar sorprendente en primera instancia si se recuerda el resultado de Diebold y Mariano (1995), donde se mostraba el resultado contrario, tanto ante la presencia de autocorrelación como de correlación contemporánea. Como ya se anticipó en el punto 5 del apartado 4.1, esta aparente contradicción no está fundamentada en el empleo de g discreta en vez de la función continua $g = e_{t,i}^2$ usada en aquel artículo, sino en otros detalles de la implementación del test. Diebold y Mariano (1995) calcularon el estadístico S_1 de su test DM usando $L(T) = 1$ aunque no existiera autocorrelación. Procediendo así, DM resulta robusto al valor de θ, aunque lo es en el sentido de que el tamaño es sesgado independientemente de θ (y el sesgo es aproximadamente de la misma cuantía en todos los casos), no en el sentido de que el tamaño es aproximadamente correcto para todo θ. Igualmente, el resto de los contrastes que estamos evaluando también serían robustos si aplicáramos el procedimiento de Bonferroni sistemáticamente, incluso cuando $\theta = 0$. Hemos llevado a cabo experimentos de los que se deriva nuestra afirmación, aunque no mostraremos aquí las tablas asociadas, para no excedernos más en la aportación de información numérica. Sin embargo, es obvio que, conocido el orden de la autocorrelación, la implementación adecuada para los tests no es ésa, sino la nuestra. De un modo idéntico a nosotros proceden Harvey et al (1997) y Dell'Aquila y Ronchetti (2004), que emplean DM con $g = e_{t,i}^2$ y, efectivamente, muestran que el tamaño de DM sí está afectado en longitudes muestrales cortas por el nivel de autocorrelación para función de pérdidas continua.[38]

9. Aparantemente, los tests no son completamente robustos a la presencia de correlación contemporánea entre las pérdidas $z_{t,1}$ y $z_{t,2}$. En muestras muy cortas ($T \leq 16$), el tamaño de los tests decrece notablemente si la correlación cruzada es muy alta (Tabla 5, $\rho = 0{,}9$),[39] llevando a contrastes que inicialmente tenían tamaño correcto (Mult2, Mult2-aprx) o próximo al correcto (DM) a separarse considerablemente del teórico. Como puede comprobarse en la sección 3, la derivación de ninguno de los seis tests evaluados descansa en supuestos sobre este tipo de correlación, de modo que este resultado no era esperado. La

[37] Nuevamente, Signos y Wilcoxon parecen quedar exentos de este problema, pero solo es debido a la aleatorización. Recuérdese el punto 2 de este apartado.

[38] Por ejemplo, en Harvey et al (1997) se presentan las siguientes cifras de tamaño empírico para DM, bajo pérdida cuadrática y nivel de significación $\alpha = 0{,}10$, para longitudes muestrales $T = 8, 16, 32, 64, 128$: si $\theta = 0$, las estimaciones de tamaño son 16,7, 13,5, 11,6, 10,9 y 10,3, respectivamente; en cambio, si $\theta = -0{,}5$, las estimaciones resultan 30,0, 20,3, 15,1, 12,4 y 11,5. Para valores de T superiores, los tamaños estimados pasan a ser aproximadamente los mismos para ambos valores de θ.

[39] Nuevamente, Signos y Wilcoxon parecen quedar exentos de este problema, pero solo es debido a la aleatorización.

explicación radica en el tipo de función de pérdidas g empleada. De hecho, todos los trabajos que evalúan DM bajo función cuadrática $g = e_{t,i}^2$ confirman su robustez a la correlación contemporánea (Diebold y Mariano (1995), Harvey et al (1997) y Dell'Aquila y Ronchetti (2004)).

Lo que sucede en nuestro caso es que, cuando ρ es muy alto, la probabilidad teórica $p_s = P(ZZ = 0)$ aumenta considerablemente, mientras descienden el resto de probabilidades p_i, $i \neq s$. En el Cuadro 3 a continuación se adjuntan los valores teóricos de estas probabilidades asociados al diseño de los ejercicios.[40] La circunstancia citada es precisamente aquella bajo la que ya habíamos advertido a lo largo del documento que las propiedades de los tests se degradan. En el tercer punto del apartado 3.2.2 se anticipó que el estimador de la función de probabilidad usado por Mult2, que emplea la expresión de la función de probabilidad exacta pero bajo el estimador $\overline{p}$ en vez de p, parece ser más impreciso cuando p_s es alto, pese a que posiblemente no aumente la traza de la matriz de varianzas y covarianzas de $\overline{p}$. Por su parte, en el apartado 3.2.3 mostrábamos que la distribución asintótica empleada por el test Mult2-aprx como distribución de contraste del estadístico $b\widehat{p}$ deja de constituir una buena aproximación a la distribución exacta cuando p_s es elevado, como puede verse en el tercer gráfico de la Figura 1 en dicho apartado. A esto se suma el hecho de que la varianza del estimador $\widehat{W_{\overline{p}}}$ de la varianza W_p del estadístico de contraste $\sqrt{T}b\widehat{p}$, tiende a aumentar cuanto mayor es p_s. Daremos pruebas empíricas de esto último más adelante. Finalmente, para DM, la explicación es la misma que para Mult2-aprx, dada la estrecha conexión entre ambos tests, al menos en el caso $h = 1$ (véase 3.2.3).

Lógicamente, estos problemas solo atañen a longitudes muestrales cortas, ya que tanto $\overline{p}$ como $\widehat{W_{\overline{p}}}$ son estimadores consistentes de p y W_p, respectivamente, bajo la hipótesis nula.

Cuadro 3. Valor teórico de p en los Experimentos $^{(*)}$

Ejercicios Tamaño			
	$\rho = 0,0$	$\rho = 0,5$	$\rho = 0,9$
$p_1 = P(ZZ = -2) = P(ZZ = +2) = p_5$	0,108	0,093	0,043
$p_2 = P(ZZ = -1) = P(ZZ = +1) = p_4$	0,133	0,096	0,043
$p_3 = P(ZZ = 0)$	0,518	0,622	0,828
Ejercicios Potencia			
	$\rho = 0,0$	$\rho = 0,5$	$\rho = 0,9$
$p_1 = P(ZZ = -2)$	0,193	0,178	0,135
$p_2 = P(ZZ = -1)$	0,237	0,189	0,136
$p_3 = P(ZZ = 0)$	0,388	0,510	0,705
$p_4 = P(ZZ = +1)$	0,113	0,067	0,012
$p_5 = P(ZZ = +2)$	0,069	0,056	0,012

$^{(*)}$: los valores de p son independientes del valor del parámetro θ.

En la sección próxima, se verá cómo este mismo argumento citado aquí explica discrepancias importantes entre el tamaño empírico de los contrastes y el teórico, que se producen en contextos que nada tienen que ver con una correlación contemporánea alta (de hecho, se producirán en contextos de $\rho = 0$) pero que generan probabilidades p similares a las que se mostraron en el cuadro anterior para el caso $\rho = 0{,}9$. Por tanto, la conclusión no es que los tests que evaluamos no son robustos, en general, a la correlación entre las pérdidas, sino que la no robustez del tamaño de los tests está relacionada con cuáles son las probabilidades teóricas p, y la situación $\rho = 0{,}9$ es uno de los casos que genera valores suficientemente peculiares en p como para afectar a las propiedades de cualquiera de los contrastes. En la sección 5.2 se realiza un estudio más detallado de este asunto, y se extraen conclusiones al respecto.

10. Finalmente, queremos dejar constancia de que *el tamaño del test DM no se modifica por el hecho de utilizar una función de pérdidas discreta*, en vez de la habitual función cuadrática. Para ello, basta comparar los resultados ofrecidos aquí para dicho contraste con los presentados por Diebold y Mariano (1995) y Harvey et al (1997) para el error de previsión cuadrático, en sus Tablas 4 y 1, respectivamente.

[40] Fijémonos que, conocida la función de distribución de los errores de previsión (Normal Bivariante con coeficiente de correlación igual a ρ) y las definiciones de f y g, es inmediato obtener los valores teóricos de p.

Tabla 5. Tamaño Empírico. g discreta. $\theta = 0$ ($h = 1$)
$e_{t,i} \sim N(0,1)$, $\alpha = 10\,\%$, 10000 repeticiones

	T	Signos	Wilcoxon	DM	RV-p	Mult2 (exact)	Mult2 (aprx)
$\rho = 0,0$	8	9,8	8,7	14,0	10,0	9,8	9,6
	16	10,0	8,2	13,1	15,6	10,1	10,0
	24	10,1	8,5	12,1	14,5	10,0	9,8
	32	9,9	8,6	11,7	12,8	10,0	10,0
	40	10,4	8,9	11,6	12,1	10,6	10,4
	48	10,4	9,2	11,6	11,3	10,4	10,3
$\rho = 0,5$	8	9,4	9,3	11,7	6,7	8,4	8,1
	16	10,1	8,5	13,4	14,3	10,5	10,3
	24	10,2	8,7	13,1	15,4	10,7	10,4
	32	9,9	8,2	11,3	13,9	10,0	9,8
	40	9,7	8,7	11,0	12,2	9,8	9,7
	48	9,8	8,4	10,8	11,8	9,6	9,5
$\rho = 0,9$	8	7,6	7,2	2,8	0,9	3,5	2,0
	16	9,6	8,9	10,0	4,9	7,5	7,1
	24	9,6	9,2	12,2	10,2	9,4	9,3
	32	10,5	9,0	12,8	13,1	10,3	10,2
	40	10,2	8,8	12,4	15,0	10,3	10,0
	48	10,1	8,9	12,6	15,0	10,8	10,6

T: long. muestra; ρ: coef. correl. $e_{t,1}$,$e_{t,2}$; θ: coef. MA(1) $e_{t,i}$; h: horizonte prev.

Tabla 6. Tamaño Empírico. g discreta. $\theta = -0,5$ ($h = 2$)
$e_{t,i} \sim N(0,1)$, $\alpha = 10\,\%$, 10000 repeticiones

	T	Signos	Wilcoxon	DM	RV-p	Mult2 (exact)	Mult2 (aprx)
$\rho = 0,0$	8	12,5	12,3	28,8	14,4	13,7	14,0
	16	12,5	9,3	19,5	14,6	13,6	13,7
	24	10,4	9,2	16,4	12,5	9,0	7,9
	32	10,2	8,6	15,1	15,1	9,7	9,6
	49	10,7	8,7	14,2	16,4	10,2	10,3
	48	10,8	8,7	13,0	15,3	9,8	10,3
$\rho = 0,5$	8	11,7	11,3	28,6	10,5	10,9	11,0
	16	12,8	8,8	20,2	15,3	12,8	12,9
	24	9,8	9,4	15,8	8,3	6,9	5,5
	32	9,3	8,5	14,4	11,9	8,3	7,6
	40	9,7	8,3	13,3	14,4	9,1	8,9
	48	10,3	8,5	13,4	15,7	9,6	9,6

T: long. muestra; ρ: coef. correl. $e_{t,1}$,$e_{t,2}$; θ: coef. MA(1) $e_{t,i}$; h: horizonte prev.

Tabla 7. Potencia Empírica Ajust-Tamaño. g discreta. $\theta = 0$ ($h = 1$)
$e_{t,i} \sim N(0, \sigma_i^2)$, $\sigma_1^2 = 3$, $\sigma_2^2 = 1$, $\alpha = 10\,\%$, 10000 repeticiones

	T	Signos	Wilcoxon	DM	RV-p	Mult2 (exact)	Mult2 (aprx)
$\rho = 0,0$	8	34,1	35,2	34,6	25,4	40,1	34,6
	16	49,6	49,3	49,9	29,5	55,6	49,9
	24	61,2	60,3	60,6	35,1	65,9	60,6
	32	69,9	68,8	70,2	43,7	73,6	70,2
	40	77,7	75,8	76,3	53,8	80,2	76,3
	48	83,2	81,2	82,5	61,8	85,2	82,5
$\rho = 0,5$	8	37,5	37,2	37,5	38,1	42,4	37,5
	16	55,1	53,1	52,7	34,5	57,9	52,7
	24	68,0	64,6	64,8	41,4	70,1	64,8
	32	77,2	73,4	75,9	50,2	79,4	75,9
	40	84,5	80,4	82,4	61,1	85,5	82,4
	48	89,3	85,4	87,8	69,6	90,3	87,8
$\rho = 0,9$	8	44,4	44,6	60,6	49,3	60,4	60,6
	16	74,5	73,8	75,4	70,8	80,1	75,4
	24	88,3	86,6	87,2	77,5	90,0	87,2
	32	94,8	92,8	93,4	83,2	95,2	93,4
	40	97,5	96,2	96,8	89,9	97,7	96,8
	48	98,8	97,8	98,4	93,7	98,7	98,4

T: long. muestra; ρ: coef. correl. $e_{t,1}$,$e_{t,2}$; θ: coef. MA(1) $e_{t,i}$; h: horizonte prev.

Tabla 8. Potencia Empírica Ajust-Tamaño. g discreta. $\theta = -0,5$ ($h = 2$)
$e_{t,i} \sim N(0, 1)$, $\sigma_1^2 = 3$, $\sigma_2^2 = 1$, $\alpha = 10\,\%$, 10000 repeticiones

	T	Signos	Wilcoxon	DM	RV-p	Mult2 (exact)	Mult2 (aprx)
$\rho = 0,0$	8	34,1	35,5	28,8	21,1	35,8	31,6
	16	47,5	45,9	43,7	25,1	50,1	45,8
	24	46,9	46,6	55,1	27,7	54,0	46,7
	32	55,7	54,5	63,2	33,8	62,1	55,2
	40	63,0	61,1	69,2	36,2	68,4	61,5
	48	69,7	66,4	76,3	40,7	73,9	67,8
$\rho = 0,5$	8	37,6	37,8	31,2	24,6	39,4	33,7
	16	52,5	49,7	47,9	30,0	54,9	49,7
	24	52,3	50,8	61,3	41,2	60,0	53,6
	32	62,7	59,8	70,5	40,5	68,4	61,9
	40	71,1	66,8	77,3	47,6	75,1	69,2
	48	77,3	72,6	81,9	51,5	79,8	75,1

T: long. muestra; ρ: coef. correl. $e_{t,1}$,$e_{t,2}$; θ: coef. MA(1) $e_{t,i}$; h: horizonte prev.

Cuadro 4. Autocorrelac. muestrales orden 1 serie zz_t

	$\theta = 0,0$			$\theta = -0,5$		
T	$\rho = 0,0$	$\rho = 0,5$	$\rho = 0,9$	$\rho = 0,0$	$\rho = 0,5$	$\rho = 0,9$
8	-0,12	-0,11	-0,11	-0,05	-0,06	-0,09
16	-0,06	-0,06	-0,06	0,02	0,01	-0,03
24	-0,04	-0,04	-0,04	0,05	0,03	-0,01
32	-0,03	-0,03	-0,03	0,07	0,05	0,01
40	-0,02	-0,03	-0,02	0,08	0,06	0,01
48	-0,02	-0,02	-0,02	0,08	0,07	0,02

Promedios de las autocorrelaciones muestrales de los experimentos.

Cuadro 5. Correlac. muestral
entre series $z_{t,1}$ y $z_{t,2}$.
Caso $\theta = 0$[41]

T	$\rho = 0,0$	$\rho = 0,5$	$\rho = 0,9$
8	-0,0046	0,1400	0,5387
16	-0,0040	0,1568	0,5699
24	-0,0031	0,1641	0,5863
32	-0,0024	0,1657	0,5953
40	-0,0016	0,1671	0,6011
48	-0,0013	0,1689	0,6043

Promedio correlac muestral expmtos

[41] Las estimaciones que se obtienen para el coeficiente de correlación contemporánea entre las pérdidas en el caso $\theta = -0,5$ son prácticamente iguales, por lo que presentamos solo las correspondientes al caso $\theta = 0$.

5. Algunos análisis de robustez

5.1. Análisis de robustez a la elección de la función de pérdida discreta

Los resultados presentados en la sección anterior se obtuvieron para una definición concreta de la función discreta de pérdidas g. Trataremos ahora de comprobar la robustez de las propiedades de los tests a la especificación de g. Recordemos que, en nuestro contexto, definir g significa (a) definir una partición para el dominio de datos y previsiones o de errores de previsión y (b) asignar valores a cada uno de los cuadrantes o regiones especificadas en (a). Hasta ahora, la función utilizada en las simulaciones particionaba el dominio de errores en 3 regiones, según éstos fueran $> +1$ (G+), < -1 (G-) ó en el intervalo $[-1, +1]$ (P), y las evaluaba asimétricamente respecto al signo (pérdida 2 para G-, pérdida 1 para G+ y pérdida 0 para P). En esta sección, repetiremos los ejercicios de Monte Carlo de la sección 4 bajo distintas definiciones de g (pero solo para un escenario concreto de correlación cruzada y autocorrelación). En primer lugar, se modificará la especificación de la partición variando la anchura de la zona central P, mientras se mantiene la misma asignación de pérdidas que se utilizó en la sección previa. En segundo lugar, se diseña un ejercicio complementario a éste, manteniendo fija la partición, y variando la asignación de pérdidas. En concreto, se probarán, además de la función asimétrica de 3 pérdidas distintas, funciones simétricas de 2 y 3 pérdidas (ie, que valoran los errores según magnitud, independientemente del signo), respectivamente, y, finalmente, una función asimétrica pero con 4 pérdidas distintas.

El tamaño de los contrastes no debería estar influido significativamente por el diseño elegido para g y esto es lo que se tratará de comprobar en los nuevos experimentos. En cambio, sería razonable encontrar cierta sensibilidad en la potencia a cambios en la función de pérdidas, en la medida que éstos impliquen diferentes "grados de falsedad" de la hipótesis nula.

Por su parte, la función f para la comparación de las pérdidas $z_{t,1}$ y $z_{t,2}$ no es objeto de estudio en esta sección. En todos los casos utilizaremos la misma función empleada hasta ahora, $f(z_{t,1}, z_{t,2}) = z_{t,2} - z_{t,1}$.

5.1.1. Estudio de Monte Carlo

Diseño del Experimento Para el objetivo que se ha descrito arriba, basta analizar la sensibilidad de los resultados de tamaño y potencia de los tests bajo uno cualquiera de los escenarios posibles presentados en el apartado 4.1. Por simplicidad, elegiremos el escenario $\theta = \rho = 0$. Por otro lado, dado que a estas alturas del artículo pensamos que queda claro que, de entre los tests evaluados, RV-p y Signos pueden descartarse como tests adecuados para contrastar igualdad de capacidad predictiva en el contexto de funciones g discretas,[42] reduciremos el conjunto de tests a aplicar en los experimentos a los otros cuatro. Hemos incluido el test de Wilcoxon, aunque nuestra predisposición a su uso es baja, dados los malos resultados mostrados en la Tabla 19 del Apéndice E (véase el comentario 2 del apartado 4.2). Teniendo en cuenta estas restricciones, se repiten los ejercicios de tamaño y potencia de la sección 4 pero bajo las siguientes definiciones de g:

1) Experimento A: Para la partición habitual del dominio de errores de previsión en 3 regiones (G+, G- y P), se prueban cinco anchuras para la zona central P. Es decir, en todos los casos G- se define por $e \in (-\infty, -l)$, G+ por $e \in (+l, +\infty)$ y P por $e \in (-l, +l)$, pero se adoptan cinco valores alternativos de l, a saber: 0,25, 0,50, 1, 1,50 y 2. Por su parte, la asignación de pérdidas es la misma que se ha utilizado en los ejercicios de la sección 4, una función asimétrica según signo, con 3 pérdidas distintas, que son: G-: 2, G+: 1 y P: 0. El objetivo del experimento es comprobar a) si el tamaño de los tests es robusto al modo en que se especifique la partición, y b) si la potencia está afectada por este factor, y en qué sentido.

2) Experimento B: Se comparan cuatro definiciones de g alternativas, para chequear si los contrastes son robustos a la forma en que se asignan las pérdidas en las particiones. Las tres primeras (llamémoslas g_1, g_2 y g_3) emplean ahora una partición del dominio de errores de 5 regiones: G+ y G- siguen siendo las mismas que en la sección 4, pero ahora la zona central pasa a dividirse en 3, M- (error mediano negativo), M+ (error mediano positivo) y P. Las funciones g_2 y g_3 valoran simétricamente los signos de los errores, de modo que solo consideran relevante el valor absoluto de éstos, mientras g_1 penaliza más los errores

[42] La razón para descartar el test de Signos como candidato para contrastar igualdad de capacidad predictiva radica en su escasa potencia en experimentos en los que la probabilidad de obtener pérdidas de distinto signo es igual, pero, en cambio, la probabilidad de que éstas sean de magnitud grande o pequeña depende del signo de las mismas. Véase el comentario 3 del apartado 4.2.

negativos que los positivos. Puede comprobarse que g_1 no es sino la función empleada en Experimento A y en los ejercicios de la sección 4. Por otro lado, g_1 y g_2 podrían haberse especificado con particiones de 3 y 2 regiones, respectivamente, pero empleamos 5 para facilitar la comparación con g_3.

Finalmente, g_4 utiliza una partición de 4 regiones y asigna pérdidas de forma asimétrica, según signo del error (nuevamente más graves los negativos). Esta función es similar a la función que se ha empleado para Experimento A y los ejercicios de la sección 4, solo que divide la zona central de aquella partición en dos, para poder asignar hasta 4 pérdidas distintas. En el Cuadro 6 se presenta la definición exacta de las funciones g_1, g_2, g_3 y g_4.

El objetivo del experimento es comprobar a) si el tamaño de los tests es robusto al número de pérdidas empleadas y a su asignación en la partición, y b) si la potencia está afectada por este factor, y en qué sentido.

Hay dos detalles relevantes sobre el diseño del Experimento B para el caso en el que la función de pérdidas es g_4:

a) En primer lugar, aunque la definición de la función f de comparación de pérdidas se mantiene como $f(z_{t,1}, z_{t,2}) = z_{t,2} - z_{t,1}$, esto supone que el número de valores posibles a tomar por dicha función pasará de $K = 5$ a $K = 7$. Es decir, f se representa por (F3) en vez de por (F2), que es la función que resulta en el resto de casos.

b) En segundo lugar, debido a lo anterior, el coste computacional de calcular Mult2 aumenta considerablemente. Para mitigar este problema, aplicaremos los ejercicios de Monte Carlo para g_4 solo respecto a las longitudes muestrales más pequeñas, $T = 8, 16, 24$. Esto es suficiente para el análisis de robustez que se pretende.

$$
\begin{array}{cc|cccc}
 & & \multicolumn{4}{c}{z_{t,2}} \\
 & & 0 & 1 & 2 & 3 \\
\hline
 & 0 & 0 & +1 & +2 & +3 \\
z_{t,1} & 1 & -1 & 0 & +1 & +2 \\
 & 2 & -2 & -1 & 0 & +1 \\
 & 3 & -3 & -2 & -1 & 0
\end{array}
\tag{F3}
$$

Cuadro 6. Funciones de pérdida Experimento B

	G-	M-	P	M+	G+
	$(-\infty, -1)$	$(-1, -0{,}5)$	$(-0{,}5, +0{,}5)$	$(+0{,}5, +1)$	$(+1, +\infty)$
g_1	2	0	0	0	1
g_2	1	0	0	0	1
g_3	2	1	0	1	2

	G-	P-	P+	G+
	$(-\infty, -1)$	$(-1, 0)$	$(0, +1)$	$(+1, +\infty)$
g_4	3	1	0	2

Resultados De los experimentos realizados se extraen las siguientes conclusiones:

1. El tamaño de los contrastes es, en general, robusto a la definición de la función de pérdidas g. De todas las situaciones diseñadas en los experimentos, solamente en los casos de las particiones que emplean $l \geq 1{,}50$ (Experimento A) el tamaño de ambos tests se deteriora claramente, resultando menor que el teórico, aunque esto únicamente afecta a muestras muy cortas: en el caso $l = 1{,}50$, cuando $T = 8$; en el caso $l = 2$, cuando $T \leq 24$. La explicación vuelve a ser la misma que se argumentó en el apartado 4.2 respecto al sesgo del tamaño en escenarios de correlación contemporánea alta entre las pérdidas (Tabla 5, caso $\rho = 0{,}9$): las distintas estimaciones realizadas por los contrastes se vuelven imprecisas cuando la probabilidad $p_s = P(ZZ = 0)$ toma valores muy elevados. Ésta es precisamente la situación cuando se define g con una región central demasiado grande, igual que ocurría cuando la correlación contemporánea entre pérdidas era muy alta aunque entonces la zona central de la partición tuviera una delimitación "adecuada". En el Cuadro 7 se adjuntan los vectores p teóricos asociados a cada uno de los nueve casos explorados en los Experimentos A y B, para los ejercicios de tamaño. Los correspondientes a las regiones con $l = 1{,}50$ y con $l = 2$ del Experimento A muestran una estructura similar a la identificada en el diseño

$\rho = 0{,}9$ del ejercicio de tamaño de la sección 4: valor numérico alto para p_s, y valores pequeños y similares entre sí para el resto de p_i. En el apartado 5.2 se presentará un breve estudio que pretende caracterizar la forma que debe tener el vector teórico p para que el tamaño de los tests se degrade a la baja cuando la muestra es muy corta, como ocurre en las situaciones que hemos mencionado.

Por su parte, la asignación de pérdidas no resultó relevante en ningún caso respecto al tamaño de los tests. Ni la simetría o asimetría en la valoración del signo de los errores de previsión ni el número de pérdidas distintas consideradas parecen jugar ningún papel en el tamaño de los tests, en general.[43] Esto ocurre simplemente porque ninguna de estas modificaciones respecto a la función usada en la sección 4 (función g_1) altera la estructura de p del modo mencionado arriba.

Cuadro 7. Valor teórico de p Experimentos A y B. Ejercicios de Tamaño

	Experimento A					Experimento B			
	l					g			
	$0,25$	$0,50$	$1,00$	$1,50$	$2,00$	g_1	g_2	g_3	g_4
$P(ZZ=0)$	0,36	0,33	0,52	0,76	0,91	0,52	0,56	0,34	0,28
$P(ZZ=-1)=P(ZZ=+1)$	0,24	0,21	0,13	0,06	0,02	0,13	0,22	0,21	0,20
$P(ZZ=-2)=P(ZZ=+2)$	0,08	0,11	0,11	0,06	0,02	0,11	-	0,12	0,11
$P(ZZ=-3)=P(ZZ=+3)$	-	-	-	-	-	-	-	-	0,05

2. Por el contrario, la potencia de los tests sí es sensible a la definición de la partición asociada a la función de pérdidas (Tabla 9b). Conforme más ancha es la zona central P de la partición, mayor es la potencia de los contrastes. En principio, la explicación podría ser simplemente que la "falsedad" de la hipótesis nula es mayor en dichos casos, pero esto no es exacto. La hipótesis nula de los cuatro contrastes es $bp = 0$,[44] mientras el verdadero valor de bp en todos los ejercicios de potencia es negativo. En la penúltima fila del Cuadro 8 se presenta el valor de bp en cada ejercicio. En principio, sería esperable que, cuanto más negativo sea bp −por tanto, más falsa la hipótesis nula−, mayor sea la potencia de los tests. Sin embargo, esta afirmación no se cumple con precisión. Por ejemplo, para $l = 1$ y $l = 1{,}5$ el valor de bp es prácticamente el mismo, y, sin embargo, los tests son más potentes en el segundo caso. Aún más, los tests son considerablemente más potentes para $l = 2$ que para $l = 1$, aun cuando el verdadero valor de bp es menos negativo en el primer caso. La explicación es que cada valor de l tiene asociado un valor numérico del vector paramétrico p diferente, que puede observarse también en el Cuadro 8, y éste no solo define la posición central de la distribución teórica del estadístico de contraste, posición caracterizada por su esperanza matemática, sino que también influye en su dispersión. En la última fila del Cuadro 8 puede encontrarse el valor numérico de la varianza teórica del estadístico de contraste $b\widehat{p}$, usado por Mult2, Mult2-aprx y DM,[45] para cada uno de los vectores p de estos ejercicios, calculado para longitud muestral $T = 8$.[46] Sean dos vectores p cuyas distribuciones del estadístico de contraste asociadas tienen la misma esperanza matemática, pero diferente varianza. Obviamente, el test será más potente en el caso de la distribución con menor varianza. Esto es lo que ocurre en nuestro contexto. Por ejemplo, la dispersión de la verdadera distribución en el caso $l = 2$ es mucho menor que en el caso $l = 1$, lo que explica que, pese a que la posición de la distribución se haya modificado menos respecto a la hipótesis nula que en $l = 1$, la potencia de los contrastes sea ahora mayor.

En cambio, igual que ocurrió en los ejercicios para tamaño, la asignación de pérdidas no resulta influyente en la potencia de los tests. Con los datos de media y varianza del estadístico $b\widehat{p}$ suministrados en el Cuadro 8, puede comprobarse que los cambios en los valores numéricos de la función de pérdidas g tienden a modificar la posición y dispersión de la distribución del estadístico de forma paralela, lo que produce pocas alteraciones en la potencia del test. Es decir, a diferencia de lo que ocurrió ante cambios

[43] La única excepción la protagoniza el test de Wilcoxon, cuando se usa la función de pérdidas g_2 (simétrica, solo dos valores de pérdida). En tal caso, el test infraestima el tamaño, convergiendo éste a 5 %, cuando el teórico era 10 %.

[44] La hipótesis nula del test DM es $E(d_t) = 0$, pero en nuestro contexto de pérdidas discretas, se tiene que $E(d_t) = bp$. Por su parte, la hipótesis del contraste de Wilcoxon es $med(d_t) = 0$, pero como estamos utilizando distribuciones simétricas, se verifica $med(d_t) = E(d_t)$, por lo que la hipótesis nula vuelve a ser $bp = 0$.

[45] En el Apéndice 3 se comprobó que, en el contexto de funciones de pérdida discretas, el estadístico $\overline{d}$ del test DM es exactamente el estadístico $b\widehat{p}$.

[46] La expresión de $V(b\widehat{p})$ es $V(b\widehat{p}) = T^{-1}\left(\sum b_i^2 p_i - (bp)^2\right)$.

en l, ahora aquellas funciones que generan mayores aumentos en el valor absoluto de bp también producen mayores incrementos de la varianza de la distribución, por lo que la masa de probabilidad a la izquierda del punto crítico (potencia) tiende a variar poco. Por eso, aunque en base al valor de bp, se esperaría, por ejemplo, que la potencia de los tests fuera mayor usando g_4 que bajo g_2, no ocurre así, porque la varianza de la distribución también es muy superior para g_4 que para g_2 (la media, en valor absoluto, es el doble, pero la varianza −en el caso $T = 8$− es cinco veces mayor).

Cuadro 8. Valor teórico de p, bp y $V(\widehat{bp})$ (ésta, en $T = 8$). Ejercicios de Potencia.

	Experimento A					Experimento B			
	l					g			
	0,25	0,50	1,00	1,50	2,00	g_1	g_2	g_3	g_4
$P(f=-3)$	-	-	-	-	-	-	-	-	0,10
$P(f=-2)$	0,09	0,15	0,19	0,17	0,12	0,19	-	0,22	0,19
$P(f=-1)$	0,26	0,27	0,24	0,18	0,12	0,24	0,38	0,25	0,21
$P(f=0)$	0,38	0,32	0,39	0,56	0,72	0,39	0,48	0,33	0,24
$P(f=+1)$	0,22	0,19	0,11	0,05	0,02	0,11	0,14	0,13	0,15
$P(f=+2)$	0,05	0,07	0,07	0,04	0,02	0,07	-	0,07	0,07
$P(f=+3)$	-	-	-	-	-	-	-	-	0,04
bp	-0,12	-0,23	-0,37	-0,38	-0,30	-0,37	-0,25	-0,40	-0,49
$V(\widehat{bp})$ en $T=8$	0,128	0,160	0,157	0,115	0,076	0,157	0,058	0,170	0,304

Tabla 9a. Experimento A. Tamaño Empírico, según anchura región $[-l, +l]$ en función g.
$e_{t,i} \sim N(0,1)$, $\alpha = 10\,\%$, 5000 repeticiones [$\theta = 0$ ($h = 1$), $\rho = 0$]

T	$l = 0,25$	$l = 0,5$	$l = 1,0$	$l = 1,5$	$l = 2,0$	$l = 0,25$	$l = 0,5$	$l = 1,0$	$l = 1,5$	$l = 2,0$
	Test Wilcoxon					Test DM				
8	7,6	6,9	8,9	8,6	4,9	14,5	14,5	13,7	5,7	0,5
16	7,9	7,9	8,2	9,0	7,7	12,8	12,7	13,1	13,0	3,2
24	8,0	8,1	8,3	8,3	8,9	12,2	12,0	12,2	12,7	7,0
32	8,2	8,7	8,3	8,4	8,4	11,6	11,8	11,2	11,8	9,0
40	8,1	8,9	8,9	8,1	8,5	11,4	11,4	11,5	11,5	11,0
48	8,0	9,2	9,1	8,0	9,3	11,1	11,3	11,5	11,1	12,1
	Test Mult2					Test Mult2-aprx				
8	9,9	9,6	9,2	4,9	3,1	10,5	10,0	9,5	3,6	0,3
16	9,9	9,6	9,8	10,1	4,1	9,8	9,4	9,8	9,6	2,1
24	9,8	9,6	10,2	10,4	5,4	9,7	9,5	9,7	10,4	4,8
32	10,2	10,2	9,6	10,0	7,5	10,1	10,0	9,5	10,0	6,4
40	10,2	10,2	10,4	10,2	8,7	10,1	10,1	10,2	9,9	8,1
48	9,9	10,5	10,2	9,8	9,2	9,5	10,3	10,0	9,5	9,1

g particiona el dominio de $e_{t,i}$ en: $(-\infty, -l)$, $[-l, +l]$, $(+l, +\infty)$, asignando pérdidas 2, 0 y 1, respectivamente.

Tabla 9b. Experimento A. SAP Empírica, según anchura región $[-l, +l]$ en función g.
$e_{t,i} \sim N(0,1)$, $\alpha = 10\,\%$, 5000 repeticiones [$\theta = 0$ ($h = 1$), $\rho = 0$]

T	$l = 0,25$	$l = 0,5$	$l = 1,0$	$l = 1,5$	$l = 2,0$	$l = 0,25$	$l = 0,5$	$l = 1,0$	$l = 1,5$	$l = 2,0$
	Test Wilcoxon					Test DM				
8	17,2	23,2	33,2	40,3	35,5	17,4	23,7	33,8	40,9	48,3
16	20,4	31,1	49,1	59,3	58,0	21,2	32,5	49,8	56,5	74,1
24	23,7	37,7	59,3	71,1	71,6	24,2	38,0	62,6	73,9	74,5
32	27,3	44,0	68,6	79,1	81,0	29,2	46,2	72,4	82,4	83,3
40	29,7	49,5	75,4	85,2	87,1	32,5	51,6	78,9	88,5	88,9
48	31,7	53,4	80,7	90,5	92,0	34,2	56,1	83,4	92,7	93,2
	Test Mult2					Test Mult2-aprx				
8	16,0	24,1	40,4	54,5	59,8	17,4	23,7	33,8	40,9	48,3
16	19,5	30,8	55,1	69,8	79,1	21,2	32,5	49,8	56,5	74,1
24	23,3	37,2	67,6	82,6	88,4	24,2	38,0	62,6	73,9	74,5
32	27,9	45,4	75,7	88,8	93,8	29,2	46,2	72,4	82,4	83,3
40	31,0	50,4	81,6	93,0	96,3	32,5	51,6	78,9	88,5	88,9
48	32,5	55,3	85,7	95,8	98,1	34,2	56,1	83,4	92,7	93,2

g particiona el dominio de $e_{t,i}$ en: $(-\infty, -l)$, $[-l, +l]$, $(+l, +\infty)$, asignando pérdidas 2, 0 y 1, respectivamente.

Tabla 10a. Experimento B. Tamaño Empírico, según funciones pérdida.
$e_{t,i} \sim N(0, \sigma_i^2)$, $\sigma_1^2 = 3$, $\sigma_2^2 = 1$, $\alpha = 10\,\%$, 5000 repeticiones [$\theta = 0$ ($h = 1$), $\rho = 0$]

T	Func g_1	Func g_2	Func g_3	Func g_4	Func g_1	Func g_2	Func g_3	Func g_4
	Test Wilcoxon				Test DM			
8	9,4	8,5	8,0	8,7	14,8	12,7	15,4	15,8
16	8,4	6,6	8,1	9,4	12,8	12,5	12,8	14,0
24	8,4	6,2	8,6	9,5	12,0	11,9	12,3	12,5
32	8,3	5,4	8,3	-	10,7	11,7	11,4	-
40	8,1	5,4	8,3	-	10,3	11,4	10,8	-
48	8,1	5,1	8,6	-	11,1	12,1	10,9	-
	Test Mult2				Test Mult2-aprx			
8	10,4	10,0	10,1	10,3	10,2	11,4	10,4	10,2
16	10,3	9,7	9,9	10,9	10,0	9,5	9,7	10,6
24	10,1	9,7	10,0	10,5	9,9	10,4	10,1	10,4
32	9,5	9,6	9,6	-	9,4	9,8	9,7	-
40	9,4	9,4	9,6	-	9,3	9,3	9,6	-
48	9,9	10,2	9,9	-	9,6	10,0	9,9	-

g_1, g_4: asimétricas 3 y 4 pérdidas; g_2, g_3: simétricas 2 y 3 pérdidas. Con g_4, se impone $T \leq 24$.

Tabla 10b. Experimento B. SAP Empírica, según funciones pérdida.
$e_{t,i} \sim N(0, \sigma_i^2)$, $\sigma_1^2 = 3$, $\sigma_2^2 = 1$, $\alpha = 10\,\%$, 5000 repeticiones [$\theta = 0$ ($h = 1$), $\rho = 0$]

T	Func g_1	Func g_2	Func g_3	Func g_4	Func g_1	Func g_2	Func g_3	Func g_4
	Test Wilcoxon				Test DM			
8	33,4	35,4	33,0	33,9	33,9	28,6	33,8	33,6
16	48,7	53,0	48,5	47,8	49,6	52,1	51,6	49,5
24	59,3	65,9	60,9	59,2	61,0	66,7	64,2	60,0
32	68,4	74,7	71,5	-	71,4	75,2	74,7	-
40	75,7	79,8	78,5	-	79,2	82,5	81,4	-
48	81,0	83,5	83,9	-	84,1	87,4	86,1	-
	Test Mult2				Test Mult2-aprx			
8	39,3	41,2	36,4	37,8	33,9	28,6	33,8	33,6
16	54,6	58,8	53,6	52,0	49,6	52,1	51,6	49,5
24	65,9	69,6	65,8	63,0	61,0	66,7	64,2	60,0
32	75,8	79,6	75,4	-	71,4	75,2	74,7	-
40	82,3	85,2	81,8	-	79,2	82,5	81,4	-
48	86,5	89,5	86,7	-	84,1	87,4	86,1	-

g_1, g_4: asimétricas 3 y 4 pérdidas; g_2, g_3: simétricas 2 y 3 pérdidas. Con g_4, se impone $T \leq 24$.

5.2. Análisis de robustez al valor del vector p teórico

A lo largo del documento se ha mostrado que los contrastes Mult2 (ambas versiones) y DM presentan un tamaño bastante ajustado al teórico, sobre todo el primero. Sin embargo, en las dos secciones anteriores también se identificaron situaciones puntuales en las que los tests infraestimaban significativamente el tamaño si las muestras eran muy cortas. En concreto, se detectaron dos casos en los que sucedía esto: cuando la correlación contemporánea entre las pérdidas de los dos conjuntos era muy alta (sección 4) y cuando la partición asociada a la definición de la función de pérdidas g empleaba una región central excesivamente ancha, de modo que los errores de previsión se situaban casi siempre dentro de ella (sección 5). Resulta que el vector de probabilidades p tenía el mismo "aspecto" en ambos casos: el parámetro central p_s elevado, mientras el resto de parámetros p_i tomaban valores pequeños y similares entre sí. Pues bien, lo que pretendemos ahora es caracterizar de forma más precisa cómo deben ser los valores del vector teórico p para que el tamaño de los tests presente sesgo. Otra cuestión distinta es determinar o conocer de antemano qué tipo de escenarios predictivos pueden dar lugar a tales valores.

5.2.1. Diseño del estudio

El análisis que realizaremos será muy sencillo. Calcularemos el tamaño de los tests Mult2, Mult2-aprx y DM en $T = 8$ para una gran variedad de valores prefijados de p y seleccionaremos aquéllos para los que el tamaño resultante esté por debajo de un umbral u_I o por encima de otro umbral u_S. Llevaremos a cabo el mismo análisis para longitud muestral $T = 16$ pero, en ese caso, restringiendo solo a los contrastes Mult2-aprx y DM, para evitar asumir el aumento de coste computacional de Mult2 que se produce al incrementarse T, y teniendo en cuenta que, de cara al objetivo del estudio, la información de estos dos tests será suficiente. Lógicamente, elegimos longitudes muestrales muy cortas porque es en ellas donde el problema de sesgo en tamaño resulta significativo, y usamos dos ($T = 8$ y $T = 16$) para comprobar cuánto se mitiga la distorsión al aumentar ligeramente el número de observaciones disponibles.

En este tipo de estudio ya no es necesario diseñar un escenario de previsión, en términos de distribución generadora de errores de previsión, autocorrelación y correlación contemporánea entre ellos. Fijémonos que los tres tests requieren para su aplicación solamente el vector b de posibles valores a tomar por la función f de comparación entre pérdidas y el vector n^0 con las frecuencias observadas para los K valores de b, siendo n^0 realización de una distribución Multinomial de probabilidades teóricas p. Para que el lector quede convencido de esta afirmación, repásense los apartados que presentan los tests Mult2 y Mult2-aprx, así como la propiedad [Prop 3] de 3.2.3, que relacionaba el test DM con Mult2-aprx en el caso $h = 1$. Debe quedar claro que, en un contexto de funciones de pérdida discretas, todo el escenario de los experimentos de simulación del tipo a los de las secciones previas está caracterizado completamente por el valor numérico del vector p.

Además, en este estudio se calculará el tamaño t^* de los contrastes *de forma exacta*, en vez de estimarse. El procedimiento es el siguiente:

a) Para cada posible vector de frecuencias n^0 cuyos elementos sumen T, se aplica el test. Si se rechaza la hipótesis nula, se asignará $I = 1$ y, si no, $I = 0$.

b) Se añade el resultado que se obtuviera para n^0 al cálculo de t^* del siguiente modo $t^* = t^* + I \cdot P_M(n_1^0, n_2^0, ..., n_K^0|p)$, siendo $P_M(n_1^0, n_2^0, ..., n_K^0|p)$ la función de masa de una Multinomial de dimensión K para el vector de frecuencias observadas n^0 y vector de probabilidades teóricas p.

Esta forma de proceder es muy exhaustiva, ya que, para un nivel de T dado, se ejecuta el contraste para todos y cada uno de los posibles vectores de frecuencias cuyos componentes sumen exactamente T, es decir, se ejecuta $\frac{T+K-1!}{T!K-1!}$ veces. Ese número crece enormemente con T y, si $K = 5$, es mayor que 10000 (el número habitual de repeticiones para los ejercicios de las secciones anteriores) para cualquier $T \geq 20$. Ésta es la razón por la que en los ejercicios presentados en secciones anteriores el tamaño se estimó, utilizando simulación, en vez de calcularse de forma exacta.

Para limitar el número de vectores p que se van a probar en el estudio, se utilizan las siguientes restricciones:

(i) Los elementos de p deben ser de la forma $p_i = \frac{j}{100}$, siendo j un número entero entre 1 y $100-K+1$.[47]

[47] No se pueden admitir componentes nulos en p, ya que, de existir, la probabilidad $P_M(n_1^0, n_2^0, ..., n_K^0|p)$ no podría calcularse. De no haber sido por esto, j hubiera sido un entero entre 0 y 100.

(ii) Como se trata de calcular el tamaño de los tests, la hipótesis nula $bp = 0$ debe ser cierta. Por ello, y aunque la condición a continuación no es necesaria sino solo suficiente para cumplir $bp = 0$, forzaremos a que p verifique $p(i) = p(K - i + 1)$, para $i = 1, ..., s - 1$, siendo $s = \frac{K+1}{2}$. Es decir, la probabilidad del suceso $ZZ = b_i$ deberá ser igual que la probabilidad del suceso $ZZ = -b_i$.

Utilizaremos $\alpha = 0{,}10$, $u_I = 0{,}09$ y $u_S = 0{,}12$, mientras supondremos que la función f puede tomar los valores -2, -1, 0, +1 y +2, es decir, $K = 5$ y $b = (-2, -1, 0, +1, +2)$. Para $K = 5$, existen 1176 vectores p que verifiquen las restricciones (i)-(ii). Para cada uno de ellos, se ejecutará el ejercicio expuesto, seleccionándose aquellos para los que el tamaño de los tests esté fuera de los umbrales propuestos.

5.2.2. Resultados

1. En primer lugar, el ejercicio sirve para confirmar los resultados de la sección 4, al extender los escenarios allí diseñados (al fin y al cabo, cada uno de los tres diseños se correspondía con un valor de p, y ahora se evalúan 1176). En general, Mult2 es el test más exacto en tamaño en longitudes muestrales muy cortas ($T = 8$), seguido de cerca por su versión asintótica, Mult2-aprx. Siempre y cuando se verifique $p_s = P(f = 0) \leq 0{,}6$, estos dos contrastes no tienen ningún sesgo en tamaño, como puede apreciarse en el resumen ofrecido en el Cuadro 9. Por el contrario, para esos mismos valores de p_s, DM sobreestima tamaño, con un sesgo casi siempre próximo al 5 %.

2. De los 1176 vectores evaluados para longitud $T = 8$, ocurrió $t^* < 0{,}09$ (sesgo por defecto) en 177, 251 y 104 de ellos, en los tests Mult2, Mult2-aprx y DM, respectivamente. Prácticamente todos estos casos se corresponden con vectores donde $p_s > 0{,}6$. La infraestimación del tamaño todavía no es grave en vectores donde $0{,}6 < p_s \leq 0{,}72$, pero sí lo es a partir de dicho valor. En el caso más extremo, cuando $p_s > 0{,}90$, los tests carecen por completo de fiabilidad: el tamaño de Mult2-aprx y DM prácticamente es cero (no rechazan la hipótesis nula casi nunca), mientras el de Mult2 apenas supera 0,03, cuando α era 0,10. Por otro lado, los contrastes Mult2 y Mult2-aprx no presentaron un tamaño por encima de 0,12 (sesgo por exceso) en ninguna de las 1176 pruebas, mientras, por el contrario, DM lo hace sistemáticamente en vectores con $p_s \leq 0{,}6$ (en 983 de los 1176 vectores), si bien su sesgo no suele sobrepasar el 5 %. Solo en 89 de los 1176 casos consultados, el tamaño de DM se mantuvo en el rango [0,09, 0,12]. Todos estos resultados pueden consultarse en el Cuadro 10.

Además, hemos observado que el único patrón explicativo del sesgo en tamaño es el valor de p_s, mientras, por el contrario, el hecho de que los valores del resto de probabilidades p_i sean o no similares entre sí, no es relevante a estos efectos. Recordemos que los vectores p que habían producido sesgo en tamaño en ejercicios de simulación de secciones anteriores, presentaban, además de un p_s elevado, también valores no muy diferentes en el resto de p_i. Sin embargo, ahora hemos comprobado que tal estructura fue puramente casual.[48]

3. Cuando la longitud muestral alcanza los 16 datos, los resultados de los tests mejoran. Como se observa en Cuadro 11, siempre que $p_s \leq 0{,}8$, Mult2-aprx presenta tamaños exactos, mientras el sesgo de DM se reduce a 3 %. Aún cuando $0{,}8 < p_s \leq 0{,}92$, los sesgos de los tests son aceptables, y solo en casos de $p_s > 0{,}9$, la infraestimación del tamaño sigue siendo muy grave.[49] Nuevamente, Mult2-aprx no genera sesgo por exceso ($t^* > 0{,}12$) nunca, mientras DM sigue haciéndolo prácticamente siempre que $p_s \leq 0{,}8$ (Cuadro 12).

4. Como conclusión fundamental del ejercicio, podemos decir que, definiendo la función de pérdidas discreta g y la de comparación f de manera que $K = 5$, los tests de naturaleza discreta evaluados en este documento presentan problemas de infraestimación grave del tamaño en escenarios predictivos donde $p_s > 0{,}7$ si $T = 8$, pero la distorsión solo se produce cuando $p_s > 0{,}9$ si $T = 16$. Es decir, el conflicto se produce en situaciones donde la probabilidad de que las dos previsiones a comparar tengan asignada la misma pérdida sea muy elevada. Como hemos venido argumentando a lo largo del documento, la causa debe ser un aumento notable en tales casos de la imprecisión de ciertos estimadores utilizados por los tests: el estimador de la función de probabilidad $P(\overline{ZZ} = w)$ en Mult2,[50] que viene dado por (7), y el

[48] No exponemos datos al respecto para no complicar innecesariamente la exposición.

[49] Por supuesto, Mult2 obtendría aún mejores resultados, pero recuérdese que, por evitar el coste computacional de su cálculo, se ha restringido el análisis a los otros dos tests.

[50] Tal estimador consiste simplemente en aplicar la expresión (5) pero sustituyendo el verdadero vector paramétrico p por su estimador MVR bajo la hipótesis nula, $\overline{p}$.

estimador de W_p en los tests Mult2-aprx y DM.[51] El parámetro p_s toma valores muy altos, por ejemplo, cuando los conjuntos de previsiones están muy correlados, o cuando la definición de la partición de la función de pérdidas es muy poco fina (una zona central de tal amplitud que las previsiones (o errores de previsión) se sitúan fuera de ella muy infrecuentemente). Tal y como se puede inferir de los resultados de las simulaciones de la sección 4 (Tabla 5, caso $\rho = 0{,}9$), el problema del sesgo en tamaño desaparecerá en cualquier caso si la muestra disponible alcanzara las 24 observaciones, independientemente del valor de p, ya que los estimadores anteriormente citados son consistentes.

5. A lo largo del documento, hemos venido apuntando que el sesgo que aparece en el tamaño de los tests en situaciones donde p_s es alto y donde, además, el tamaño de la muestra es pequeño, viene motivado por los incrementos de la varianza del estimador $\widehat{P}(\overline{ZZ} = w) = \zeta_w(\overline{p})$ de Mult2 y de la varianza del estimador $\widehat{W_{\overline{p}}}$ de Mult2-aprx (análogamente, ocurre con el estimador $\widehat{W_{\widehat{p}}}$ de DM), respectivamente. Demostrar esta afirmación para el primero de estos estimadores es muy complicado (podría hacerse por simulación), dada la dificultad de encontrar la expresión analítica de la varianza $V(\zeta_w(\overline{p}))$. Sin embargo, la varianza del segundo es conocida y su expresión se mostró en la propiedad [Prop 2b] de 3.2.3. Así que podemos calcular los valores de $V(\widehat{W_{\overline{p}}})$ para distintos valores numéricos del vector p, y comprobar si la varianza relativa $V(\widehat{W_{\overline{p}}})/W_p$ se incrementa cuando p_s es alto, tal y como argumentábamos. Esto es lo que se presenta en el Cuadro 13. Las cifras allí mostradas sugieren que nuestra explicación es correcta.

Cuadro 9. Resumen Tamaño Tests $\alpha = 10\,\%$
1176 valores de p. $T = 8$

Rango $p_s = P(ZZ = 0)^{(*)}$	Tamaño medio (%)			Sesgo medio (%) (val. abs)		
	Mult2 (exact)	Mult2 (aprx)	DM	Mult2 (exact)	Mult2 (aprx)	DM
0,02; 0,10	10,3	9,6	14,9	0,5	0,6	4,9
0,12; 0,20	10,3	9,9	14,9	0,3	0,3	4,9
0,22; 0,30	10,2	9,9	14,9	0,2	0,4	4,9
0,32; 0,40	10,2	10,3	14,7	0,2	0,4	4,7
0,42; 0,50	10,1	10,6	14,4	0,2	0,6	4,4
0,52; 0,60	9,5	10,0	13,2	0,5	0,8	3,2
0,62; 0,70	8,0	8,1	10,5	2,0	2,0	1,0
0,72; 0,80	6,0	5,2	6,7	4,0	4,8	3,3
0,82; 0,90	4,0	2,3	3,0	6,0	7,7	7,0
0,92; 0,96	3,3	0,4	0,6	6,7	9,6	9,4

$^{(*)}$: Se refiere a todos los vectores cuyo elemento central se encuentra en ese rango

Cuadro 10. Información sobre vectores p que generaron sesgo. $T = 8$

		Sesgo Defecto ($t^* < 9\,\%$)						Sesgo Exceso ($t^* > 12\,\%$)					
		nº vect $t^* < 9\,\%^{(2)}$			t^* medio (%)$^{(3)}$			nº vect $t^* > 12\,\%^{(2)}$			t^* medio (%)$^{(3)}$		
Rango p_s	nº vect$^{(1)}$	Mult2		DM	Mult2		DM	Mult2		DM	Mult2		DM
		exact	aprx		exact	aprx		exact	aprx		exact	aprx	
0,02; 0,10	230	0	49	1	-	8,5	8,7	0	0	208	-	-	15,3
0,12; 0,20	205	0	16	0	-	8,8	-	0	0	205	-	-	14,9
0,22; 0,30	180	0	17	0	-	8,7	-	0	0	180	-	-	14,9
0,32; 0,40	155	0	2	0	-	8,9	-	0	0	155	-	-	14,7
0,42; 0,50	130	0	0	0	-	-	-	0	0	130	-	-	14,4
0,52; 0,60	105	11	16	0	8,9	8,7	-	0	0	100	-	-	13,1
0,62; 0,70	80	75	60	12	7,9	7,5	8,8	0	0	5	-	-	12,1
0,72; 0,80	55	55	55	55	6,0	5,2	6,7	0	0	0	-	-	-
0,82; 0,90	30	30	30	30	4,0	2,3	3,0	0	0	0	-	-	-
0,92; 0,96	6	6	6	6	3,3	0,4	0,6	0	0	0	-	-	-
Totales	1176	177	251	104				0	0	983			

$^{(1)}$ nº vect p probados; $^{(2)}$ nº vect p que generaron sesgo; $^{(3)}$ media incluyendo solo casos de sesgo; t^*: tamaño test.

[51] En el caso de Mult2-aprx y DM, además, influye el hecho de que la distribución asintótica es una muy mala aproximación de la verdadera distribución del estadístico de contraste cuando p_s es muy grande (ver Figura 1, en 3.2.3).

Cuadro 11. Resumen Tamaño Tests $\alpha = 10\,\%$
1176 valores de p. $T = 16$

	Tamaño medio (%)		Sesgo medio (%) (val. abs)	
Rango p_s $= P(ZZ = 0)$	Mult2 (aprx)	DM	Mult2 (aprx)	DM
$0,02; 0,10$	9,9	12,9	0,2	2,9
$0,12; 0,20$	9,9	13,0	0,1	3,0
$0,22; 0,30$	9,9	13,0	0,1	3,0
$0,32; 0,40$	10,1	12,9	0,1	2,9
$0,42; 0,50$	10,1	13,0	0,1	3,0
$0,52; 0,60$	10,1	13,2	0,1	3,2
$0,62; 0,70$	10,3	13,3	0,3	3,3
$0,72; 0,80$	9,8	12,5	0,6	2,5
$0,82; 0,90$	6,8	9,1	2,9	1,5
$0,92; 0,96$	1,2	1,6	8,8	8,4

Cuadro 12. Información sobre vectores p que generaron sesgo. $T = 16$

		Sesgo Defecto ($t^* < 9\,\%$)				Sesgo Exceso ($t^* > 12\,\%$)			
		nº vects $t^* < 9\,\%$		t^* medio (%)		nº vects $t^* > 12\,\%$		t^* medio (%)	
Rango p_s	nº vects	Mult2 (aprx)	DM	Mult2 (aprx)	DM	Mult2 (aprx)	DM	Mult2 (aprx)	DM
$0,02; 0,10$	230	4	0	8,8	-	0	216	-	13,0
$0,12; 0,20$	205	0	0	-	-	0	205	-	13,0
$0,22; 0,30$	180	0	0	-	-	0	177	-	13,0
$0,32; 0,40$	155	0	0	-	-	0	151	-	13,0
$0,42; 0,50$	130	0	0	-	-	0	130	-	13,0
$0,52; 0,60$	105	0	0	-	-	0	105	-	13,2
$0,62; 0,70$	80	0	0	-	-	0	80	-	13,3
$0,72; 0,80$	55	12	0	8,6	-	0	38	-	12,6
$0,82; 0,90$	30	29	19	6,7	7,1	0	0	-	-
$0,92; 0,96$	6	6	6	1,2	1,6	0	0	-	-
Totales	1176	51	25			0	1102		

Cuadro 13. Varianza estimador $\widehat{W_{\overline{p}}}$ (Mult2-aprx)
según p_s; $b = (-2, -1, 0, +1, +2)$

		Ratio $V(\widehat{W_{\overline{p}}})/W_p$		
Valor p_s $= P(ZZ = 0)$	p_i, $\forall i \neq s$	$T = 1$	$T = 8$	$T = 16$
$0,40$	$0,150$	1,900	0,237	0,119
$0,50$	$0,125$	2,150	0,269	0,134
$0,60$	$0,100$	2,400	0,300	0,150
$0,70$	$0,075$	2,650	0,331	0,165
$0,80$	$0,050$	2,900	0,362	0,181
$0,90$	$0,025$	3,150	0,394	0,197
$0,95$	$0,0125$	3,275	0,410	0,205

6. Conclusiones

Hemos comprobado que las propiedades del contraste de Diebold y Mariano (1995) (DM) no varían si éste se implementa bajo función de pérdida discreta, respecto a las que obtuvieron los autores bajo pérdida cuadrática. La excepción se produce en el caso de que las muestras de errores de previsión contengan algún valor atípico. En dichas situaciones, los resultados del test DM implementado con error cuadrático se distorsionan drásticamente, mientras el contraste es robusto a la presencia de atípicos si se aplicó con una función de pérdidas discreta.

Además de esto, hemos evaluado un grupo de seis tests que contrastan igualdad de capacidad predictiva entre dos conjuntos de previsiones, centrándonos exclusivamente en un contexto de funciones de pérdida discretas. Tres de los tests son habituales en la literatura, entre ellos el contraste DM, mientras los otros son presentados en este documento como novedad. Hemos realizado experimentos de Monte Carlo, utilizando especificaciones de diseño estándar en la literatura, para chequear las propiedades de los tests en muestras finitas, y se obtiene evidencia de que los contrastes DM y Mult2 presentan los mejores resultados. La potencia de ambos es muy similar pero Mult2 mantiene un tamaño más exacto en todas las simulaciones realizadas, bajo distintos escenarios de autocorrelación y correlación cruzada entre las pérdidas asociadas a las previsiones de los dos conjuntos, y bajo varias definiciones de la función de pérdidas. La diferencia en exactitud del tamaño entre ambos tests (a favor de Mult2) es notable en casos de autocorrelación en las pérdidas y muestras con un número de previsiones inferior a 20. Solo en ciertos casos peculiares, que tratamos de caracterizar al final del trabajo, el tamaño del test puede presentar sesgo en muestras muy pequeñas (no más de 16 datos), algo que también ocurre con el resto de los contrastes. Mult2 emplea una estimación de la distribución exacta de su estadístico de contraste, en vez de la distribución asintótica, hecho que explica sus mejores propiedades en muestras finitas en relación a DM, en el contexto de funciones de pérdidas discretas. Mult2 puede computarse siempre, mientras DM podría plantear problemas de negatividad de la varianza del estadístico $\overline{d}$ en algunas ocasiones, cuando las pérdidas estén autocorrelacionadas. La contrapartida de Mult2 es su alto costo computacional si el número de valores posibles de la función f para la comparación de pérdidas (K) es alto o si la longitud de la muestra de previsiones es grande. Sin embargo, ofrecemos también la versión asintótica de Mult2 (Mult2-aprx), que soluciona el problema anterior, con pérdidas razonablemente pequeñas en potencia y nulas en tamaño. Nuestra recomendación para situaciones predictivas donde procede aplicar la función discreta es utilizar Mult2 si la muestra no contiene más de 50 datos y $K \leq 5$ ó si no contiene más de 25 datos y $K = 7$ y, en caso contrario, emplear el habitual test DM o Mult2-aprx.[52] Todo esto, suponiendo que no exista autocorrelación en la muestra de pérdidas. Bajo autocorrelación de orden r, las restricciones sobre el número de datos de la muestra de previsiones se relajan: puede usarse Mult2 para longitudes muestrales $r + 1$ veces superiores a las que acabamos de exponer. En general, Mult2 es especialmente recomendable (frente a DM) en casos de autocorrelación en las pérdidas y longitud muestral inferior a 20.

[52] Si el usuario no deseara programar Mult2, para la elección entre DM y Mult2-aprx recomendamos Mult2-aprx si existe autocorrelación en los diferenciales de pérdidas y la muestra es corta, digamos $T < 20$.

A. Apéndice: Implementación de los contrastes

En este Apéndice trataremos de especificar cómo se ha llevado a cabo exactamente la aplicación de los contrastes en los experimentos de Monte Carlo del trabajo, tratando de resumir cuestiones ya mencionadas durante el documento y exponiendo alguna otra que no se hubiera detallado hasta ahora.

A.1. Detalles del cálculo de los estadísticos de contraste

• Test DM: Se utilizará siempre "lag-window" rectangular y $L(T) = h-1$ para la estimación de $f_d(0)$. De este modo, la expresión para S_1 queda $S_1 = \frac{\overline{d}}{\sqrt{\widehat{V}(\overline{d})}}$, con $\widehat{V}(\overline{d}) = T^{-1}\left(\widehat{\gamma}_d(0) + 2\sum_{\tau=1}^{h-1}\widehat{\gamma}_d(\tau)\right)$, donde $\widehat{\gamma}_d(\tau)$, sigue la expresión habitual, ya presentada en el apartado 3.1.1.

• Test de Signos: Tras eliminar todos los elementos $d_t = 0$ de la secuencia original $\{d_t\}_{t=1}^{T}$, se obtiene una nueva serie de longitud T'. Sobre ésta, se utiliza la versión asintótica del estadístico (S_{2a}) si $T' \geq 64$ y la de muestras finitas (S_2) en caso contrario. El límite $T' = 64$ es el mismo usado por Diebold y Mariano (1995) en sus simulaciones.

• Test de Wilcoxon: Se obtiene primero la serie de longitud T', igual que en Signos. Se utiliza la versión asintótica del estadístico (S_{3a}) si $T' > 20$.[53]

Respecto al cálculo de $rank(|d_t|)$ en caso de "empates" se aplica la regla siguiente: sean m elementos iguales de la secuencia $\{d_t\}_{t=1}^{T'}$ a los que teóricamente corresponde rango r, entonces se aplica a todos ellos $rank(|d_t|) = \frac{r+(r+1)+...+(r+m-1)}{m}$.

• Test Mult2: Se sustituyen las estimaciones $\overline{p}_i = 0$ por $\overline{p}'_i = \xi$, para garantizar que el test sea computable. Se usa $\xi = 0,0001$ para minimizar la alteración de los datos muestrales. Se utiliza el método de aleatorización explicado en el segundo apartado de 3.2.5.

• Test Mult2-aprx: Se sustituyen las estimaciones $\overline{p}_i = 0$ por $\overline{p}'_i = \xi$ en aquellas muestras en que la varianza $\widehat{W}_{\overline{p}}$ resultó nula. Se usa $\xi = 0,0001$. Esto garantiza no rechazar la hipótesis nula en muestras en las que $zz_t = 0\ \forall t$, único caso en el que ocurre $\widehat{W}_{\overline{p}} = 0$.

• Test RV-p: Se sustituyen las estimaciones $\widehat{p}_i = 0$ por $\widehat{p}'_i = \xi$, para garantizar que el test sea computable. Se usa $\xi = 0,0001$.

A.2. Casos de imposibilidad de computar los tests

Nuestra implementación de los tests garantiza que sean computable para cualquier muestra, salvo el test DM en casos de autocorrelación de d_t:

a) Mult2-aprx es computable por construcción, excepto en un único caso: cuando todos los elementos de $\widehat{p}$ son nulos salvo el central, que toma valor igual a uno. Este caso se resuelve aplicando el método descrito en el primer apartado de 3.2.5 y mencionado de nuevo en A.1, lo que conduce directamente a no rechazar la hipótesis nula, tal y como es deseable.

b) RV-p y Mult2: en teoría, serían no computables en aquellos casos en los que $\widehat{p}_i = 0$ y $\overline{p}_i = 0$, respectivamente. Pero en el primer apartado de 3.2.5 se explicaron mecanismos sencillos para que los contrastes Mult2 y RV-p tuvieran garantizada su aplicación. Aconsejamos fijar el valor del parámetro ξ que se empleaba en dichos procedimientos en $\xi = 0,0001$.

c) Signos y Wilcoxon: en teoría, no serían computables cuando $d_t = 0\ \forall t$, ya que, tras eliminar observaciones nulas de la muestra, se tendría otra de longitud $T' = 0$. En el tercer apartado de 3.2.5, se expuso la corrección para salvar este obstáculo, pasando directamente a no rechazarse la hipótesis nula $H_0^{(SW)}$.

d) DM: este contraste no puede computarse si $\widehat{V}(\overline{d}) \leq 0$, varianza cuya expresión puede consultarse en el apartado A.1. Teniendo en cuenta que siempre se implementan con "lag-window" rectangular y $L(T) = h-1$ y que utilizamos la corrección especificada en el tercer apartado de 3.2.5, esta situación solo es posible si $h > 1$.

[53] El límite $T' = 20$ se ha escogido para poder utilizar las tablas del estadístico de Wilcoxon que habitualmente ofrecen los manuales de estadística y evitar el coste computacional de generar la función de distribución para valores de T superiores a 20.

En las Tablas 11-14 del Apéndice B se presentan estimaciones de la frecuencia en que la ejecución de los tests no sería posible de no emplearse los mecanismos de corrección que acabamos de citar. En el Cuadro 2 de la sección 4 se adjuntó el porcentaje de veces que DM no pudo computarse en casos $h = 2$.

A.3. Cálculo de la Potencia Ajustada a Tamaño (SAP)

A.3.1. Planteamiento

Si se evalúan las propiedades estadísticas de un conjunto de tests en los que al menos uno es sesgado en tamaño, la comparación entre las potencias estimadas está distorsionada por el hecho de que los tamaños de los contrastes difieran. Los tests con tendencia a sobreestimar tamaño habitualmente presentarán valores estimados de potencia mayores que los tests con tendencia a infraestimarlo, sin que signifique que distinguen mejor los casos de hipótesis nula falsa, sino que esto ocurre simplemente porque su probabilidad de rechazo es sistemáticamente mayor, sea la hipótesis cierta o falsa. Para corregir esta distorsión procede modificar el mecanismo de decisión de los contrastes en ejercicios destinados a evaluar la potencia del conjunto de contrastes, en el siguiente sentido:

En los ejercicios de tamaño análogos a los de potencia (es decir, mismo diseño en cuanto a escenarios de simulación, funciones de pérdida elegidas, longitudes muestrales utilizadas, etc), se registran los valores obtenidos para el estadístico de contraste de cada test en todas las repeticiones efectuadas del ejercicio, y, con ellos, se estima la función de distribución de dicho estadístico bajo la hipótesis nula, distribución que denotamos por F_0, mientras su estimación la denotamos por $\widehat{F}_0$. Para aquellos contrastes que resultaron sesgados en tamaño, la función F_0 difiere de la función de distribución teórica utilizada en la aplicación del contraste. Pues bien, la implementación del contraste en cada repetición del ejercicio de potencia será la habitual solo que utilizando $\widehat{F}_0$ en vez de la función de distribución teórica. Esto permitirá que las mediciones de potencia de los distintos tests sean homogéneas. Esa forma de medir la potencia es lo que hemos venido denominando SAP ("size-adjusted power").

Por ejemplo, supongamos que se pretende comparar dos tests cuyos estadísticos de contraste x_1 y x_2 siguen ambos una distribución χ^2_6, en teoría. Supongamos que uno de ellos resultó exacto en tamaño, según las simulaciones realizadas con nivel de significación $\alpha = 0,1$, mientras el segundo es sesgado por exceso, habiéndose estimado un tamaño empírico igual a $0,18$. El valor crítico es $\lambda_\alpha = 10,646$, es decir, $P(\chi^2_6 > 10,6446) = 0,9$. Por tanto, se está verificando que, cuando la hipótesis nula es cierta, $P(x_1 \geq 10,6446) = 0,1$, tal y como establecía teoría del contraste, pero, sin embargo, $P(x_2 \geq 10,6446) = 0,18$. De este modo, si usamos la distribución teórica χ^2_6 en la implementación de los tests en los ejercicios de potencia, y, por consiguiente, empleamos $\lambda_\alpha = 10,6446$ como valor crítico para los dos tests, se tenderá a encontrar muchos más rechazos para el segundo que para el primero. Pero esto es debido a que se utiliza un valor crítico para el que la probabilidad de rechazo es mayor para el segundo contraste que para el primero, ya cuando la hipótesis era cierta, de manera que no parten ambos en "igualdad de condiciones". La comparación no es correcta, por consiguiente.

En cambio, el procedimiento propuesto corrige la distorsión de la comparación de potencias: se estima la función de distribución empírica F_0 del estadístico x_2, se calcula el valor crítico adecuado λ'_α, tal que $\widehat{F}_0(\lambda_\alpha) \approx 0,9$ −que será mayor que λ_α−, y es éste el que se emplea en la aplicación del segundo test cuando se evalúe potencia. Ahora sí se están empleando valores críticos tales que ambos dejen a la derecha el 10 % de la masa de probabilidad cuando la hipótesis es cierta. De este modo, el criterio SAP resulta independiente del sesgo que presenten los tests en su tamaño.

A.3.2. Detalles de aplicación del procedimiento a nuestros ejercicios de simulación

a) Dado que el tamaño de los contrastes varía según cada parámetro utilizado en el diseño de los ejercicios de tamaño, estimamos la función de distribución F_0 para cada test en cada uno de los diseños del experimento de simulación. Dichos diseños son una combinación de parámetros que definen el escenario de predicción, longitud muestral, función de pérdidas, etc. En concreto:

− Ejercicios SAP de la sección 2.1 (Tablas 2 y 4): para la Tabla 2, cada diseño se define por un valor T del tamaño muestral y una función de pérdidas g (continua o discreta), mientras para la Tabla 4, cada diseño se caracteriza por T y g, y además, por la distribución utilizada para los errores de previsión ($N(0,1)$ ó $t(3)$).

– Ejercicios SAP de la sección 4 (Tablas 7 y 8) y del Apéndice B (Tablas 15 y 16): cada diseño se define por la terna (T, θ, ρ).

Existe una excepción a estas caracterizaciones, que explicamos a continuación. En los tests de Signos y Wilcoxon, hay que estimar una función F_0 para cada $n = 1, 2, ..., \text{máx}(T) = 48$. El motivo es que estos tests utilizan solo una submuestra de la muestra original, aquella que queda tras eliminar las observaciones $d_t = 0$, por lo que, para cada ejecución del test en un ejercicio de longitud muestral teórica T, puede requerirse la función $\widehat{F}_0$ asociada a cualquier longitud entre 1 y T. Obviamente, para cada valor n, no se tendrán 10000 realizaciones del estadístico de contraste, sino que dicha cantidad se reparte entre las T posibles longitudes resultantes de la muestra finalmente utilizada. Cuando el número de realizaciones disponibles para una longitud n sea inferior a 1000 (suele ocurrir para valores n próximos a T), utilizaremos la función de distribución teórica, en vez de estimarla. Dado que estos tests presentan, debido a la aleatorización, tamaños prácticamente exactos, esto no generará errores en el cálculo de SAP.

– Ejercicios SAP de la sección 5 (Tablas 9b y 10b): para la Tabla 9b, cada diseño se caracteriza por un valor de T y un valor del parámetro l que define la anchura de la partición asociada a la función de pérdidas g, mientras para la Tabla 10b, cada diseño viene definido también por T y por una función de pérdidas g.

b) En el caso de tests cuya distribución de contraste es de tipo continuo, la aplicación del procedimiento SAP en los ejercicios de potencia solo requiere calcular y utilizar el percentil λ_α tal que $\widehat{F}_0(\lambda_\alpha) = \alpha$ ó $\widehat{F}_0(\lambda_\alpha) = 1 - \alpha$ (según cuál sea la cola que constituye la región crítica). Éste es el caso de los contrastes DM, RV-p, Mult2-aprx, y las versiones asintóticas de Signos y Wilcoxon. Sin embargo, si la distribución de contraste es de tipo discreto, y dado que empleamos el mecanismo de aleatorización, se requiere calcular los puntos $(b, \widehat{F}_0(b))$ y $(d, \widehat{F}_0(d))$ tales que $\widehat{F}_0(b) < \alpha < \widehat{F}_0(d)$ y, luego, aplicar dicho mecanismo de la manera habitual: es decir, no se rechaza la hipótesis nula cuando el estadístico de contraste es mayor que d, mientras se rechaza si, o bien es menor o igual que b, o bien es igual a d y un experimento de Bernouilli de probabilidad $\widehat{p}^* = \frac{\alpha - \widehat{F}_0(b)}{\widehat{F}_0(d) - \widehat{F}_0(b)}$ resulta "éxito".[54] Éste es el caso del test Mult2, y de los tests Signos y Wilcoxon cuando se emplean sus versiones exactas.

[54] Por simplicidad, acabamos de definir los dos puntos relevantes de $\widehat{F}$ particularizando a la situación que se produce en nuestros ejercicios de potencia, donde la región crítica es la inferior. Si fuera la superior, la definición sería $\widehat{F}_0(d) < 1 - \alpha < \widehat{F}_0(b)$. En tal caso, no se rechazaría si el estadístico de contraste es menor que d y se rechazaría si es mayor o igual que b o igual a d pero con éxito en el experimento de Bernouilli.

B. Apéndice: Tablas de aplicabilidad empírica de los tests

A continuación se adjuntan Tablas (11-14) con los porcentajes de veces en que cada uno de los tests no pudieron ser computados en experimentos idénticos a los de la sección 4, solo que implementados sin aplicar ninguno de los procedimientos de corrección empleados para garantizar computabilidad, que se mencionaron en A.2. La intención es cuantificar la relevancia de usar estos mecanismos en la implementación práctica de los contrastes.

En segundo lugar, se exponen comparativas para los resultados de tamaño y potencia de los tests Mult2 (Tabla 15) y RV-p (Tabla 16) entre las simulaciones donde se resolvió el caso $\overline{p}_i$, $\widehat{p}_i = 0$ sustituyendo por $\overline{p}'_i$, $\widehat{p}'_i = 0{,}0001$ (son los resultados de las Tablas 5-8) y simulaciones idénticas pero donde el tamaño y potencia se calculan contabilizando solamente muestras para las que no se produjeron problemas de probabilidades nulas (es decir, muestras donde ocurrió $\overline{p}_i \neq 0\ \forall i$ en el caso de Mult2, y muestras donde $\widehat{p}_i \neq 0\ \forall i$ en el caso de RV-p).

Tabla 11. Frecuencia empírica (%) de no aplicabilidad.
H_0 cierta. $\theta = 0$ $(h = 1)$
$e_{t,i} \sim N(0,1)$, $\alpha = 10\,\%$, 10000 intentos

	T	Signos	Wilcoxon	DM	RV-p ($\xi = 0$)	Mult2 ($\xi = 0$)	Mult2 (aprx)
$\rho = 0,0$	8	0,5	0,5	0,5	87,5	22,6	0,5
	16	0,0	0,0	0,0	45,6	2,7	0,0
	24	0,0	0,0	0,0	18,6	0,3	0,0
	32	0,0	0,0	0,0	6,9	0,0	0,0
	40	0,0	0,0	0,0	2,8	0,0	0,0
	48	0,0	0,0	0,0	1,0	0,0	0,0
$\rho = 0,5$	8	2,2	2,2	2,2	94,1	36,0	2,2
	16	0,1	0,1	0,1	62,2	7,1	0,1
	24	0,0	0,0	0,0	32,3	1,4	0,0
	32	0,0	0,0	0,0	15,5	0,3	0,0
	40	0,0	0,0	0,0	7,0	0,0	0,0
	48	0,0	0,0	0,0	3,0	0,0	0,0
$\rho = 0,9$	8	21,6	21,6	21,6	99,6	75,4	21,6
	16	4,7	4,7	4,7	94,6	42,2	4,7
	24	1,0	1,0	1,0	83,3	22,3	1,0
	32	0,2	0,2	0,2	68,5	11,2	0,2
	40	0,0	0,0	0,0	53,5	5,5	0,0
	48	0,0	0,0	0,0	40,6	2,8	0,0

Mult2 (RV-p) no se computó si $\exists i$ tq $\overline{p}_i = 0$ ($\widehat{p}_i = 0$) (es decir, se usó $\xi = 0$).

Tabla 12. Frecuencia empírica (%) de no aplicabilidad.
H_0 cierta. $\theta = -0,5$ $(h = 2)$
$e_{t,i} \sim N(0,1)$, $\alpha = 10\,\%$, 10000 intentos

	T	Signos	Wilcoxon	DM	RV-p ($\xi = 0$)	Mult2 ($\xi = 0$)	Mult2 (aprx)
$\rho = 0,0$	8	0,1	0,1	4,4	92,9	29,5	0,1
	16	0,1	0,1	1,0	54,9	4,7	0,1
	24	0,0	0,0	0,3	88,1	14,7	0,0
	32	0,0	0,0	0,0	68,8	5,5	0,0
	40	0,0	0,0	0,0	49,1	1,9	0,0
	48	0,0	0,0	0,0	32,4	0,7	0,0
$\rho = 0,5$	8	0,1	0,1	3,1	97,3	42,1	0,1
	16	0,0	0,0	1,0	71,0	10,5	0,0
	24	0,0	0,0	0,2	95,2	28,7	0,0
	32	0,0	0,0	0,0	84,6	13,1	0,0
	40	0,0	0,0	0,0	69,6	5,8	0,0
	48	0,0	0,0	0,0	54,0	2,7	0,0

Mult2 (RV-p) no se computó si $\exists i$ tq $\overline{p}_i = 0$ ($\widehat{p}_i = 0$) (es decir, se usó $\xi = 0$).

Tabla 13. Frecuencia empírica (%) de no aplicabilidad.
H_0 falsa. $\theta = 0$ $(h = 1)$
$e_{t,i} \sim N(0,1)$, $\alpha = 10\,\%$, 10000 intentos

	T	Signos	Wilcoxon	DM	RV-p ($\xi = 0$)	Mult2 ($\xi = 0$)	Mult2 (aprx)
$\rho = 0,0$	8	0,2	0,2	0,2	85,4	13,6	0,2
	16	0,0	0,0	0,0	41,3	1,0	0,0
	24	0,0	0,0	0,0	16,7	0,1	0,0
	32	0,0	0,0	0,0	6,7	0,0	0,0
	40	0,0	0,0	0,0	2,7	0,0	0,0
	48	0,0	0,0	0,0	1,1	0,0	0,0
$\rho = 0,5$	8	1,3	1,3	1,3	92,9	21,9	1,3
	16	0,0	0,0	0,0	59,3	2,4	0,0
	24	0,0	0,0	0,0	30,1	0,2	0,0
	32	0,0	0,0	0,0	15,2	0,0	0,0
	40	0,0	0,0	0,0	7,5	0,0	0,0
	48	0,0	0,0	0,0	3,6	0,0	0,0
$\rho = 0,9$	8	16,6	16,6	16,6	99,5	49,6	16,6
	16	2,6	2,6	2,6	94,4	15,6	2,6
	24	0,5	0,5	0,5	84,7	4,7	0,5
	32	0,0	0,0	0,0	71,9	1,4	0,0
	40	0,0	0,0	0,0	60,3	0,5	0,0
	48	0,0	0,0	0,0	50,5	0,1	0,0

Mult2 (RV-p) no se computó si $\exists i$ tq $\overline{p}_i = 0$ ($\widehat{p}_i = 0$) (es decir, se usó $\xi = 0$).

Tabla 14. Frecuencia empírica (%) de no aplicabilidad.
H_0 falsa. $\theta = -0,5$ ($h = 2$)
$e_{t,i} \sim N(0,1)$, $\alpha = 10\,\%$, 10000 intentos

	T	Signos	Wilcoxon	DM	RV-p ($\xi = 0$)	Mult2 ($\xi = 0$)	Mult2 (aprx)
$\rho = 0,0$	8	0,1	0,1	5,8	91,5	20,3	0,1
	16	0,0	0,0	1,2	51,6	1,8	0,0
	24	0,0	0,0	0,3	83,8	6,8	0,0
	32	0,0	0,0	0,0	62,9	1,8	0,0
	40	0,0	0,0	0,0	44,4	0,4	0,0
	48	0,0	0,0	0,0	29,9	0,1	0,0
$\rho = 0,5$	8	0,1	0,1	5,5	95,9	26,6	0,1
	16	0,0	0,0	1,4	70,0	3,7	0,0
	24	0,0	0,0	0,2	93,5	13,1	0,0
	32	0,0	0,0	0,1	82,0	4,6	0,0
	40	0,0	0,0	0,0	66,4	1,3	0,0
	48	0,0	0,0	0,0	50,3	0,4	0,0

Mult2 (RV-p) no se computó si $\exists i$ tq $\overline{p}_i = 0$ ($\widehat{p}_i = 0$) (es decir, se usó $\xi = 0$).

Tabla 15. Comparación resultados Test Mult2, según tratamiento en casos $\overline{p}_i = 0$
$e_{t,i} \sim N(0,1)$, $\alpha = 10\,\%$, 10000 repeticiones

		TAMAÑO				Potencia Ajustada a Tamaño (SAP)			
		$\theta = 0$ ($h = 1$)		$\theta = -0,5$ ($h = 2$)		$\theta = 0$ ($h = 1$)		$\theta = -0,5$ ($h = 2$)	
	T	$\xi = 0$	$\xi = 10^{-4}$	$\xi = 0$	$\xi = 10^{-4}$	$\xi = 0$	$\xi = 10^{-4}$	$\xi = 0$	$\xi = 10^{-4}$
$\rho = 0,0$	8	9,8	9,8	12,2	13,7	41,2	40,1	38,7	35,8
	16	10,2	10,1	13,9	13,6	55,2	55,6	49,6	50,1
	24	9,8	10,0	8,8	9,0	65,1	65,9	55,8	54,0
	32	10,3	10,0	9,7	9,7	73,6	73,6	62,6	62,1
	40	9,6	10,6	9,3	10,2	80,5	80,2	69,4	68,4
	48	9,9	10,4	9,6	9,8	85,8	85,2	74,1	73,9
$\rho = 0,5$	8	9,7	8,4	10,8	10,9	44,4	42,4	45,1	39,4
	16	10,7	10,5	12,5	12,8	58,6	57,9	55,4	54,9
	24	10,1	10,7	7,1	6,9	69,8	70,1	62,4	60,0
	32	10,6	10,0	8,5	8,3	79,1	79,4	69,8	68,4
	40	10,7	9,8	9,0	9,1	84,6	85,5	76,1	75,1
	48	10,5	9,6	9,1	9,6	89,3	90,3	80,5	79,8
$\rho = 0,9$	8	5,5	3,5			85,9	60,4		
	16	9,4	7,5			84,7	80,1		
	24	10,3	9,4			91,0	90,0		
	32	10,0	10,3			95,5	95,2		
	40	10,3	10,3			97,5	97,7		
	48	9,8	10,8			99,1	98,7		

Si $\xi = 10^{-4}$, sustituimos $\overline{p}_i = 0$ por $\overline{p}'_i = \xi = 10^{-4}$; si $\xi = 0$, excluimos muestras donde $\exists i$ tal que $\overline{p}_i = 0$.

Tabla 16. Comparación resultados Test RV-p, según tratamiento en casos $\widehat{p}_i = 0$
$e_{t,i} \sim N(0,1)$, $\alpha = 10\,\%$, 10000 repeticiones

		TAMAÑO				POTENCIA			
		$\theta = 0$ $(h=1)$		$\theta = -0{,}5$ $(h=2)$		$\theta = 0$ $(h=1)$		$\theta = -0{,}5$ $(h=2)$	
	T	$\xi = 0$	$\xi = 10^{-4}$	$\xi = 0$	$\xi = 10^{-4}$	$\xi = 0$	$\xi = 10^{-4}$	$\xi = 0$	$\xi = 10^{-4}$
$\rho = 0,0$	8	0,0	10,0	0,0	14,4	0,0	25,4	0,0	21,1
	16	0,8	15,6	1,5	14,6	1,7	29,5	2,5	25,1
	24	4,4	14,5	0,0	12,5	8,7	35,1	0,0	27,7
	32	7,7	12,8	0,2	15,1	13,6	43,7	0,4	33,8
	40	10,0	12,1	0,7	16,4	17,6	53,8	2,7	36,2
	48	10,7	11,3	2,4	15,3	20,4	61,8	6,2	40,7
$\rho = 0,5$	8	0,0	6,7	0,0	10,5	0,0	38,1	0,0	24,6
	16	0,1	14,3	0,5	15,3	0,9	34,5	1,3	30,0
	24	2,1	15,4	0,0	8,3	5,7	41,4	0,0	41,2
	32	5,1	13,9	0,1	11,9	12,2	50,2	0,1	40,5
	40	7,6	12,2	0,2	14,4	17,1	61,1	0,6	47,6
	48	9,5	11,8	0,8	15,7	22,3	69,6	2,9	51,5
$\rho = 0,9$	8	0,0	0,9			0,0	49,3		
	16	0,0	4,9			0,0	70,8		
	24	0,0	10,2			0,3	77,5		
	32	0,1	13,1			2,3	83,2		
	40	0,7	15,0			6,7	89,9		
	48	1,5	15,0			14,4	93,7		

Si $\xi = 10^{-4}$, sustituimos $\overline{p}_i = 0$ por $\overline{p}_i' = \xi = 10^{-4}$; si $\xi = 0$, excluimos muestras donde $\exists i$ tal que $\overline{p}_i = 0$.

Los resultados correspondientes al caso $h = 2$ y $T = 16$ para los tests RV-p y Mult2 en las Tablas 12 y 14 resultan incoherentes aparentemente en relación a los obtenidos para longitudes muestrales superiores. Sin embargo, son correctos. La explicación reside en la implementación que decidimos emplear para los tests en longitudes muestrales menores o iguales que 16, cuando la autocorrelación no es nula ($h = 2$), casos éstos en los que no aplicamos el procedimiento de Bonferroni, a diferencia de lo que hacemos para muestras de superior tamaño. Parece extraño que el porcentaje de muestras en las que no pudieron ejecutarse los tests RV-p y Mult2 cuando no se emplea la corrección descrita en el primer apartado de 3.2.5 (representada por $\xi = 0,0001$) aumente de forma muy significativa al pasar de longitud $T = 16$ a $T = 24$, en vez de descender, tal y como uno pensaría que debe ocurrir para todo valor de T. La razón es que en $T = 16$ no se está utilizando el mecanismo de Bonferroni, mientras a partir de $T = 24$, sí. Los contrastes no se pueden aplicar en una muestra de longitud 24 cuando no pueden hacerlo en una cualquiera o las dos submuestras de longitud 12. Por tanto, la probabilidad de no aplicabilidad en $T = 24$ usando el método Bonferroni debe ser claramente mayor que en $T = 16$ si no se usa. De hecho, dicha probabilidad debe ser algo mayor en $T = 32$ que en $T = 16$, ya que basta que el test no pueda aplicarse en una de las dos submuestras de longitud 16 en que se divide la muestra de longitud 32 para que consideremos que el test no pudo aplicarse. En los valores de las Tablas 12 y 14 puede comprobarse que dicho comportamiento se verifica para los contrastes citados.

También en las Tablas 15 y 16 se observa este patrón en los resultados de los tests en los experimentos con $h = 2$. De nuevo, la monotonía respecto a T del tamaño y SAP empíricas se interrumpe entre $T = 16$ y $T = 24$. La justificación a este comportamiento se fundamenta otra vez en el asunto descrito arriba, y ya fue mencionada en una nota a pie de página en el apartado 4.2.

C. Apéndice: Comparativa entre Mult2-aprx y DM. Resultados teóricos y empíricos.

C.1. Demostraciones de las propiedades teóricas enunciadas en 3.2.3

Demostramos a continuación las propiedades que se enunciaron en el apartado 3.2.3 relacionadas con el contraste Mult2-aprx y con la implementación discreta del test DM en caso de horizonte de previsión uno:

Primero, recuérdese la definición de W_p y véase su desarrollo algebraico:

$W_p = bV_pb' = b\left(\Omega - pp'\right)b' =$

$$\left(\begin{array}{ccc} b_1 & ... & b_K \end{array} \right) \left(\begin{array}{ccc} p_1(1-p_1) & -p_1p_2 & -p_1p_K \\ -p_2p_1 & p_2(1-p_2) & -p_2p_K \\ & & \\ -p_Kp_1 & -p_Kp_2 & p_K(1-p_K) \end{array} \right) \left(\begin{array}{c} b_1 \\ . \\ . \\ b_K \end{array} \right) =$$

$$\sum_{i=1}^{K} b_i^2 p_i - \sum_{i=1}^{K} b_i p_i \left(\sum_{j=1}^{K} b_j p_j \right) = \sum_{i=1}^{K} b_i^2 p_i - (bp)^2 .$$

[Prop 1] y [Prop 2] (Estimadores de W_p)

[Prop 1] $\widehat{W_{\widehat{p}}} = \sum_{i=1}^{K} b_i^2 \widehat{p}_i - (b\widehat{p})^2$: Se deduce directamente de la expresión lograda arriba para W_p y de la definición de $\widehat{W_{\widehat{p}}}$.

[Prop 2a] $\widehat{W_{\overline{p}}} = \sum_{i=1}^{K} b_i^2 \widehat{p}_i$: Tomando la definición de $\widehat{W_{\overline{p}}}$, se tiene $\widehat{W_{\overline{p}}} = W_{p|\overline{p}} = \sum_{i=1}^{K} b_i^2 \overline{p}_i - b\overline{p}$. Como $\overline{p}_i = \overline{p}_{K-i+1}$ para $i = 1, ..., s-1$ (recuérdese que $s = \frac{K+1}{2}$ y $b_s = 0$) y, por definición del vector b, $b_i = -b_{K-i+1}$, se tiene que $b\overline{p} = b_s\overline{p}_s = 0$ y $\widehat{W_{\overline{p}}} = \sum_{i=1}^{K} b_i^2 \overline{p}_i$. Apliquemos ahora la definición del estimador $\overline{p}$: $\overline{p}_i = \frac{n_i + n_{K-i+1}}{2N} = \frac{1}{2}\left(\widehat{p}_i + \widehat{p}_{K-i+1}\right)$. Por lo tanto, $\widehat{W_{\overline{p}}} = \sum_{i=1}^{K} b_i^2 \overline{p}_i = \sum_{i=1}^{K} b_i^2 \widehat{p}_i$, q.e.d.

Respecto a la consistencia de $\widehat{W_{\overline{p}}}$: partiendo de la expresión anterior de $\widehat{W_{\overline{p}}}$, la propiedad se deduce directamente a partir de las siguientes afirmaciones: (1) $W_p = \sum_{i=1}^{K} b_i^2 p_i - bp$, (2) $H_0^{(1)} = bp = 0$, (3) $\widehat{p}$ es un estimador consistente de p y (4) del teorema sobre convergencia en probabilidad de una función continua de una sucesión de variables aleatorias convergentes en probabilidad.

Por su parte, la consistencia de $\widehat{W_{\widehat{p}}}$ se deduce directamente de (3) y (4).

[Prop 2b] $V(\widehat{W_{\overline{p}}}) = T^{-1} b^{(2)} V_p b^{(2)\prime}$, donde $b^{(2)}$ es un vector $1 \times K$ cuyo elemento iésimo es $b_i^{(2)} = b_i^2$:

$V(\widehat{W_{\overline{p}}}) = V(\sum_{i=1}^{K} b_i^2 \widehat{p}_i) = V(b^{(2)}\widehat{p}) = E[b^{(2)}(\widehat{p}-p)(\widehat{p}-p)'b^{(2)\prime}] = b^{(2)}E[(\widehat{p}-p)(\widehat{p}-p)']b^{(2)\prime} = T^{-1}b^{(2)}V_pb^{(2)\prime}$, q.e.d., donde se ha utilizado el resultado conocido respecto a la matriz de varianzas del estimador MV $\widehat{p}$, según el cual $E[(\widehat{p}-p)(\widehat{p}-p)'] = T^{-1}(\Omega - pp') = T^{-1}V_p$.

Las propiedades a continuación ([Prop 3]-[Prop 9]) se establecen todas bajo el cumplimiento del conjunto de condiciones siguiente, que denotamos por [C]:

(a) La varianza $2\pi f_d(0)$ del estadístico $\sqrt{T}\overline{d}$ de DM se estima mediante la suma de las autocovarianzas muestrales del diferencial d_t de órdenes $\tau = -(h-1), ..., h-1$, siendo h el horizonte de previsión;

(b) El horizonte de previsión es uno, $h = 1$. De (a) y (b) se deduce que la estimación de la varianza $2\pi f_d(0)$ es, simplemente, $\frac{1}{T}\sum_{t=1}^{T}(d_t - \overline{d})^2$;

(c) La función de pérdida $g(y_t, v_{t,i})$ empleada en el cálculo de DM es la misma función de pérdida discreta que en Mult2-aprx.

(d) La función de comparación de pérdidas f empleada en Mult2-aprx es $f(z_{t,1}, z_{t,2}) = zz_t = z_{t,2} - z_{t,1}$. Es decir, zz_t es el diferencial d_t del test DM.

Denotemos por S_1^D el estadístico del test DM cuando es aplicado verificando el conjunto de condiciones [C].

[Prop 3] y [Prop 4] (Relación analítica entre DM y Mult2-aprx)

[Prop 3] $S_1^D = \varphi(M2_{apx}) = \frac{M2_{apx}}{\sqrt{1-\frac{(M2_{apx})^2}{T}}}$.

[Prop 4] $\frac{M2_{apx}}{S_1^D} = \left(1 - \frac{(b\widehat{p})^2}{\sum_{i=1}^{K} b_i^2 \widehat{p}_i}\right)^{1/2}$.

– Demostremos primero que las expresiones para $M2_{apx}$ y S_1^D son:

$$M2_{apx} = \sqrt{T} b\widehat{p} \left(\sum_{i=1}^{K} b_i^2 \widehat{p}_i\right)^{-1/2} \tag{10a}$$

$$S_1^D = \sqrt{T} b\widehat{p} \left(\sum_{i=1}^{K} b_i^2 \widehat{p}_i - (b\widehat{p})^2\right)^{-1/2}. \tag{10b}$$

La expresión (10a) se deduce directamente a partir de la definición de $M2_{apx}$ y de [Prop 2].

Pasemos a demostrar (10b):

Aplicando las condiciones (a) y (b), la expresión del estadístico S_1 de DM sería $S_1^D = \frac{\overline{d}}{\sqrt{\frac{2\pi \hat{f}_{\overline{d}}(0)}{T}}} = \frac{\sqrt{T}\overline{d}}{\left(\frac{1}{T}\sum_{t=1}^{T}(d_t - \overline{d})^2\right)^{1/2}}$.

Teniendo en cuenta que $\frac{1}{T}\sum_{t=1}^{T}(d_t - \overline{d})^2 = \sum_{t=1}^{T}\frac{d_t^2}{T} - \overline{d}^2$, solo falta demostrar que $\overline{d} = b\widehat{p}$ y que $\sum_{t=1}^{T}\frac{d_t^2}{T} = \sum_{i=1}^{K} b_i^2 \widehat{p}_i$.

Pero, aplicando las condiciones (c) y (d), cada diferencial d_t es uno de los elementos del vector b, y $\sum_{t=1}^{T} d_t = \sum_{i=1}^{K} b_i n_i$ y $\sum_{t=1}^{T} d_t^2 = \sum_{i=1}^{K} b_i^2 n_i$. Dada la definición de $\widehat{p}$ ($\widehat{p}_i = \frac{n_i}{T}$), se tiene que $\overline{d} = b\widehat{p}$ y $\sum_{t=1}^{T}\frac{d_t^2}{T} = \sum_{i=1}^{K} b_i^2 \widehat{p}_i$, q.e.d, con lo que la expresión propuesta en (10b) para S_1^D queda demostrada.

– Basta ahora comparar las expresiones obtenidas para $M2_{apx}$ y S_1^D, para obtener las propiedades deseadas:

[Prop 4] $\frac{M2_{apx}}{S_1^D} = \left(\frac{S_1^D}{M2_{apx}}\right)^{-1} = \left(\left(1 - \frac{(b\widehat{p})^2}{\sum_{i=1}^{K} b_i^2 \widehat{p}_i}\right)^{-1/2}\right)^{-1} = \left(1 - \frac{(b\widehat{p})^2}{\sum_{i=1}^{K} b_i^2 \widehat{p}_i}\right)^{1/2}$.

[Prop 3] Usando [Prop 4], y teniendo en cuenta que $(M2_{apx})^2 = \frac{T(b\widehat{p})^2}{\sum_{i=1}^{K} b_i^2 \widehat{p}_i}$, se tiene que $\frac{M2_{apx}}{S_1^D} = \sqrt{1 - \frac{(M2_{apx})^2}{T}}$. Por tanto, $S_1^D = \frac{M2_{apx}}{\sqrt{1-\frac{(M2_{apx})^2}{T}}}$, q.e.d.

[Prop 5] y [Prop 6] (Corolarios a [Prop 4]):

[Prop 5] "Si Mult2-aprx usara $\widehat{W}_{\widehat{p}}$ como estimación de W_p, en vez de $\widehat{W}_{\overline{p}}$, entonces $M2_{apx} = S_1^D$".

Esta afirmación se deduce directamente de la comparación entre (10a) y (10b), y los resultados [Prop 1] y [Prop 2].

[Prop 6] "Si la hipótesis nula $H_0^{(1)}$ es cierta, $\frac{M2_{apx}}{S_1^D} \xrightarrow{p} 1$. En el resto de casos, se cumple $\frac{M2_{apx}}{S_1^D} \xrightarrow{p} \left(1 - \frac{(bp)^2}{\sum_{i=1}^{K} b_i^2 p_i}\right)^{1/2}$".

En el caso en que $H_0^{(1)}$ sea falso, el resultado se sigue directamente de [Prop 4] y de aplicar la consistencia de $\widehat{p}$ y el teorema de convergencia en probabilidad de una función continua de una sucesión de variables

aleatorias convergentes en probabilidad. Cuando $H_0^{(1)}$ es cierta, basta aplicar sobre el resultado anterior la restricción correspondiente a dicha hipótesis, $bp = 0$, para obtener $\frac{M2_{apx}}{S_1^D} \xrightarrow{p} 1$.

[Prop 7], [Prop 8] y [Prop 9] (Corolarios a [Prop 3]):

[Prop 7] "El estadístico S_1^D está definido en todos los puntos en los que lo está $M2_{apx}$ (por su parte, $M2_{apx}$ está definido para cualquier muestra salvo aquella que genere que la frecuencia relativa central sea uno, $\widehat{p}_s = 1$), salvo en los puntos $M2_{apx} = -\sqrt{T}$ y $M2_{apx} = +\sqrt{T}$ (que se corresponden con muestras que generaron alguna frecuencia relativa no central igual a uno, $\widehat{p}_i = 1$, para algún $i \neq s$)".

En primer lugar, véase que $M2_{apx}$ está definido para cualquier $\widehat{p}$, salvo si $\widehat{p}_s = 1$, siendo $b_s = 0$ (obviamente, $\widehat{p}_i = 0$, para $i \neq s$). Basta considerar la expresión (10a). En segundo lugar, a partir de [Prop 3], se deduce directamente que S_1^D no está definido en los puntos $M2_{apx} = -\sqrt{T}$ y $M2_{apx} = +\sqrt{T}$. Finalmente, usando de nuevo la expresión (10a), se deduce que $M2_{apx} = -\sqrt{T}$ se obtiene cuando $\widehat{p}_i = 1$ siendo $b_i < 0$, mientras $M2_{apx} = +\sqrt{T}$ se produce cuando $\widehat{p}_i = 1$ siendo $b_i > 0$.

[Prop 8] "En todos los puntos del soporte de $M2_{apx}$, salvo en el punto $M2_{apx} = 0$ (para el que $S_1^D = 0$), se verifica $\left|S_1^D\right| > |M2_{apx}|$".

Se deduce directamente de [Prop 3].

[Prop 9] "La función φ es monótona (creciente), y, por tanto, $P(S_1^D = \varphi(x)) = P(M2_{apx} = x)$, para cualquier punto $x \in (-\sqrt{T}, +\sqrt{T})$ del soporte de la variable aleatoria $M2_{apx}$ (siendo los extremos de dicho soporte (incluidos en él), precisamente, $-\sqrt{T}$ y $+\sqrt{T}$)".

En primer lugar, que el soporte de $M2_{apx}$ son valores reales en el rango $[-\sqrt{T}, +\sqrt{T}]$ se deduce de la expresión (10a), teniendo en cuenta que $M2_{apx}$ tiene su máximo (mínimo) en el caso $\widehat{p}_K = 1$ ($\widehat{p}_1 = 1$).

Veamos ahora que la función $y = \frac{x}{\sqrt{1-\frac{x^2}{c}}}$ es una función monótona creciente en los puntos $x \in (-\sqrt{c}, +\sqrt{c})$:

$\frac{\partial y}{\partial x} = \left(1 - \frac{x^2}{c}\right)^{-1/2} + \frac{x^2}{c}\left(1 - \frac{x^2}{c}\right)^{-3/2} = \left(1 - \frac{x^2}{c}\right)^{-1/2}\left(1 + \frac{x^2}{c-x^2}\right) = \left(1 - \frac{x^2}{c}\right)^{-1/2}\left(\frac{c}{c-x^2}\right) =$

$= \left(\frac{c}{c-x^2}\right)^{3/2} > 0$, para todo $x \in (-\sqrt{c}, +\sqrt{c})$, q.e.d.

Relación entre Mult2-aprx y DM en términos de SAP

Sean $\widehat{F}_{M2apx}$ y $\widehat{F}_{DM}$ las funciones de distribución empírica de los estadísticos $M2_{apx}$ y S_1^D, respectivamente, obtenidas en las simulaciones de los ejercicios de tamaño, y sean x_0 e y_0 los puntos que verifican $\widehat{F}_{M2apx}(x_0) = \alpha$ y $\widehat{F}_{DM}(y_0) = \alpha$, respectivamente. Los puntos x_0 e y_0 son los puntos críticos usados para no rechazar o rechazar la hipótesis nula en los ejercicios de potencia (utilizando SAP) en el caso de los tests Mult2-aprx y DM, respectivamente.[55] Debido a [Prop 9], se cumplirá la igualdad $y_0 = \varphi(x_0)$. Sean x_1 e y_1 las realizaciones de $M2_{apx}$ y S_1^D, respectivamente, en una simulación del ejercicio destinado a medir la SAP de los contrastes. Se rechazará la hipótesis nula si solo si $x_1 < x_0$ en el caso de Mult2-aprx, y si solo si $y_1 < y_0$, en DM.[56] En virtud de [Prop 9], resulta que $y_1 = \varphi(x_1) < y_0 = \varphi(x_0) \Leftrightarrow x_1 < x_0$. Por consiguiente, el valor estimado de SAP para ambos tests será idéntico. Recordemos que esta afirmación se circunscribe al cumplimiento de las condiciones [C].

C.2. Análisis de Monte Carlo para DM y Mult2-aprx en muestras largas

Adjuntamos a continuación los resultados de los mismos experimentos presentados en la sección 4, solo que efectuados solamente para los contrastes DM y Mult2-aprx y para muestras largas, hasta $T = 512$. Se pretende de este modo dar respuesta a las cuestiones A y B abiertas en 3.2.3.

[55] Se aconseja releer el apartado A.3 para comprender las afirmaciones realizadas aquí.

[56] El sentido de la desigualdad es éste en nuestros ejercicios de potencia, porque, arbitrariamente, se estableció que el segundo conjunto de previsiones fuera superior predictivamente al primero.

Tabla 17. Comparación resultados tests DM y Mult2-aprx. Muestras largas
$e_{t,i} \sim N(0,1)$, $\alpha = 10\,\%$, 10000 repeticiones

		TAMAÑO				SAP			
		$\theta = 0,0$ $(h=1)$		$\theta = -0,5$ $(h=2)$		$\theta = 0,0$ $(h=1)$		$\theta = -0,5$ $(h=2)$	
	T	DM	Mult2 (aprx)	DM	Mult2 (aprx)	DM	Mult2 (aprx)	DM	Mult2 (aprx)
$\rho = 0,0$	8	13,8	9,4	26,6	13,0	34,1	34,1	29,1	31,5
	16	12,7	9,8	19,9	13,5	49,7	49,7	43,7	46,2
	24	12,0	9,6	16,3	7,2	61,2	61,2	54,4	47,3
	32	11,0	9,6	14,8	8,8	69,4	69,4	63,4	55,2
	64	10,9	10,2	11,9	9,2	90,0	90,0	84,9	78,1
	128	10,3	9,8	11,1	9,4	99,3	99,3	97,8	95,7
	256	10,1	9,8	10,3	9,3	100,0	100,0	100,0	99,9
	512	10,2	10,2	10,3	9,6	100,0	100,0	100,0	100,0
$\rho = 0,5$	8	12,3	8,7	29,5	11,8	36,8	36,8	32,0	33,2
	16	13,2	9,9	20,9	13,7	52,7	52,7	47,9	50,3
	24	12,7	10,2	15,9	5,8	64,5	64,5	61,4	51,3
	32	12,1	10,6	14,7	7,8	75,1	75,1	70,8	61,9
	64	10,6	9,9	12,6	9,5	94,5	94,5	90,2	84,8
	128	10,5	10,1	10,9	9,4	99,7	99,7	99,1	98,0
	256	10,6	10,4	9,9	9,4	100,0	100,0	100,0	100,0
	512	9,7	9,6	9,6	9,6	100,0	100,0	100,0	100,0
$\rho = 0,9$	8	3,2	2,2			59,7	59,7		
	16	9,9	7,2			74,9	74,9		
	24	12,1	9,3			86,5	86,5		
	32	12,5	9,7			93,2	93,2		
	64	10,4	9,5			99,6	99,6		
	128	10,2	9,7			100,0	100,0		
	256	9,7	9,6			100,0	100,0		
	512	9,8	9,7			100,0	100,0		

T: tamaño muestral; ρ: coef. correl. contemp. $e_{t,1}$ y $e_{t,2}$; θ: coef. MA(1) $e_{t,i}$; h: horizonte previsión

C.3. Comparación entre DM, HLN y Mult2-aprx

En el comentario 5 del apartado 4.2 se comparaban los resultados en muestras cortas de Mult2-aprx y DM. Entonces, resaltábamos la ventaja de Mult2-aprx en casos de autocorrelación en los diferenciales de pérdidas, al evitar en buena parte el sesgo en tamaño en el que incurre DM. No obstante, en el mismo comentario, nos preguntábamos por la posibilidad de que el test HLN (corrección de DM en cierto tipo de implementaciones de éste), cuya propiedad fundamental y conocida es suavizar el sesgo en tamaño respecto a DM, obtuviera ya los buenos resultados que alcanza Mult2-aprx. Los resultados del ejercicio de simulación adecuado para responder a esta cuestión se adjuntan en la Tabla 18, a continuación.

Tabla 18. Tamaño DM, HLN, Mult2-aprx
$\theta = -0,5$ $(h=2)$; $e_{t,i} \sim N(0,1)$
$\alpha = 10\,\%$, 10000 repets

	T	DM	HLN	Mult2 (aprx)
$\rho = 0,0$	8	26,6	20,7	13,0
	16	19,9	14,9	13,5
$\rho = 0,5$	8	29,5	23,5	11,8
	16	20,9	15,0	13,7

D. Apéndice: Análisis de coste computacional del test Mult2

D.1. Análisis teórico

Tal y como se ha comentado en distintos puntos del documento, el problema de la utilización de Mult2 es su exigencia computacional, que podría definirse en términos de consumo de recursos del computador o en tiempo de ejecución, siendo ésta última la medida más habitual y la de mayor interés en este caso. Como apuntábamos en el tercer apartado de 3.2.2, la razón del elevado coste computacional de la aplicación del test es que trabaja con un número muy grande de vectores de frecuencia, que denotaremos por q. Dicho número es $q = CR(K,T) = C(T+K-1,T) = \frac{T+K-1!}{T!K-1!} = \frac{(T+K-1)(T+K-2)...(T+1)}{K-1!}$, donde CR y C designan combinaciones con y sin repetición, respectivamente.[57] Recordemos de forma sintetizada los pasos del algoritmo para la implementación de Mult2 (reléanse los dos primeros apartados de 3.2.2) que son relevantes para su coste computacional:[58]

1) Crear una matriz A que contendrá, en sus filas, todos los vectores de frecuencia tales que $n_1 + ... + n_K = T$. Será una matriz $q \times K$, que es la representación matricial del conjunto Ψ del primer apartado de 3.2.2.

2) Calcular todas las pérdidas posibles, es decir, calcular el vector $D_f = Ab'$.

3) Seleccionar de A las filas de las frecuencias adecuadas:[59] sea ZZ_n^0 el valor muestral que toma la función de comparación de pérdidas f, se escoge toda fila i de A tal que $D_f(i) \leq ZZ_n^0$. Denotemos por C el conjunto de filas seleccionadas. Su cardinalidad dependerá de cada muestra concreta y diremos que es, en general, rq, siendo $0 < r \leq 1/2$.[60]

4) Calcular el valor de la función de masa (5) para cada vector de frecuencias (fila de A) contenido en C.[61] Agregando, se tendrá $\widehat{F}(ZZ_n^0)$. Como sabemos, se rechazará la hipótesis nula si y solo si $\widehat{F}(ZZ_n^0) \leq \alpha$.

Una posible alternativa para mejorar el coste computacional del algoritmo es llevar a cabo 3) y 4) conjuntamente, seleccionando solamente las filas i de A que verifiquen $D_f(i) \leq \text{mín}(ZZ_n^0, x_m)$, siendo $\widehat{F}(x_m) > \alpha > \widehat{F}(x_{m-1})$. Es decir, seguir seleccionando filas y construyendo $\widehat{F}$ mientras no sean suficientes para no rechazar la hipótesis nula. Es posible que, sin necesidad de completar todo el conjunto C, se pueda no rechazar la hipótesis nula porque exista x_m tal que $x_m < ZZ_n^0$ y $\widehat{F}(x_m) > \alpha$, condición de no rechazo de la hipótesis. Denotaremos esta alternativa por 5).

Tal y como es habitual en el análisis de coste computacional de un algoritmo, supondremos que cada operación "simple"/"unitaria" (escribir dato en matriz, multiplicar dos enteros, calcular una potencia, etc) tiene asociado un coste constante,[62] aunque particular de cada operación. Repasemos ahora el coste computacional asociado a cada fase del algoritmo:

1) Para llevar a cabo este paso, utilizamos el siguiente algoritmo recursivo:

$soluc\_ent(K,T) : A$
si $K = 1 \rightarrow$ devolver $A = T$
si $K = 2 \rightarrow a1 = [0 : T]'$; $a2 = [T : 0]'$; devolver $A = [a1\ a2]$
si no $\rightarrow$ para $i = 0$ hasta T
$A = [A; prod\_cartes(soluc\_ent(K\ div\ 2, i), soluc\_ent(K - (K\ div\ 2), T - i))]$
fin bucle para
devolver A
fin,

[57] Véase que q debe ser el número de vectores de frecuencia $(n_1, ..., n_K)$ que suman exactamente T. Por tanto, se busca el número de grupos de T elementos que pueden formarse a partir de K tipos de elementos distintos, pudiendo repetirse o no elementos del mismo tipo (el número de veces que se elige el tipo i-ésimo es n_i). Y ésta es, precisamente, la definición en combinatoria de $CR(K,T)$.

[58] Por simplicidad, obviaremos el asunto de la aleatorización, que no introduce ningún cambio en el análisis computacional del algoritmo.

[59] Supondremos, por simplicidad, que la hipótesis alternativa es $H_1^{(1u)} \equiv bp < 0$. El coste computacional del test no depende de ello.

[60] $r \leq 1/2$ porque: a) lógicamente, solo se consideran para seleccionar elementos de D_f del mismo signo que ZZ_n^0, y b) la distribución de ZZ_n es simétrica respecto a 0. Por tanto, en el peor de los casos, habría que revisar $q/2$ elementos de D_f.

[61] Los parámetros $p_1, ..., p_K$ se habrán estimado previamente por (6).

[62] En realidad, el coste de algunas operaciones es constante mientras los operandos no sobrepasen cierto tamaño. Por ejemplo, la multiplicación de enteros grandes, de n bits, tiene un coste que depende de n (ese coste es función de n^2 o de $n\log(n)$, según el algoritmo de multiplicación empleado). Nosotros obviaremos estas cuestiones en nuestro análisis.

siendo $prod\_cartes(M_1, M_2)$ una función que realiza el producto cartesiano de las matrices M_1 y M_2, de dimensiones $m_1 \times k_1$ y $m_2 \times k_2$, devolviendo una matriz $m_1 m_2 \times k_1 k_2$ (es decir, con cada fila f_i de M_1 genera m_2 filas, siendo la j-ésima de esas filas: $[f_i \quad \text{fila } j \text{ de } M_2]$). Por otro lado, div denota el cociente entero de la división.

El coste del algoritmo $soluc\_ent(K, T)$ es el mayor de todos los pasos y el más difícil de calcular. Teniendo en cuenta que $prod\_cartes(M_1, M_2)$ tendría un coste $c_1 m_1 m_2$, y suponiendo que K es par, podemos escribir la función que expresa el coste del algoritmo como:[63]

$$t_1(K,T) = \begin{array}{l} c_1 KT, \text{ para } K \leq 2 \\ \sum_{i=0}^{T} [(c_1 i(T-i))^{K/2} + 2t_1(K/2, i)], \text{ para } K > 2. \end{array} \tag{11}$$

Esta recurrencia es de difícil resolución, y no nos detendremos en ello.

Sin embargo, la matriz A es la misma cada vez que el test Mult2 se ejecute con muestras del mismo tamaño y misma función de pérdidas, es decir, hay una matriz A para cada par (K, T). Por tanto, una vez calculada A una sola vez, lo razonable es guardarla y cargarla cada vez que se ejecuta Mult2 con los mismos valores de K y T, tal y como hacemos nosotros en todas las simulaciones que se presentan en el documento. El coste propio del algoritmo se corresponde, en realidad, con los pasos 2) a 4).

2) El coste asociado a este paso sería $t_2(q) = c_2 q$.

3) El coste asociado a este paso sería $t_3(q) = c_3 q$. Una alternativa a los pasos 2) y 3) es, si se va a ejecutar Mult2 varias veces con la misma longitud muestral y la misma función de pérdidas discreta, almacenar D_f ordenado crecientemente tras la primera ejecución, y cargar ese vector ordenado cada nueva ejecución. En tal caso, en vez de tener que chequear elemento a elemento si es mayor o menor que ZZ_n^0, puede usarse un algoritmo de búsqueda adecuado para encontrar el último elemento menor o igual que ZZ_n^0. El coste de ordenar D_f es $O(q)$ si es factible usar el algoritmo "de la casilla" y $O(q \log(q))$ si no es posible y, por el contrario, se tiene que utilizar el algoritmo habitual de "quickshort". Pero, igual que se dijo arriba respecto a t_1, este coste solo se asume una vez. Por tanto, el coste de este paso para cada ejecución del algoritmo quedaría reducido al coste de búsqueda del elemento de valor ZZ_n^0. Usando un algoritmo sencillo de "búsqueda binaria", el coste sería de orden $O(\log(q))$, agilizando notablemente la realización del paso 2).[64] La definición formal de $O(.)$ la expondremos en breve, pero puede adelantarse que si el coste de un algoritmo es de orden $O(f(n))$ viene a significar que, para n grande, el coste (tiempo de ejecución) del algoritmo nunca será superior a $kf(n)$, para alguna constante k.

En cualquier caso, nosotros no usaremos este procedimiento alternativo en el análisis computacional que vamos a exponer, ni tampoco lo recomendamos especialmente, ya que, como veremos, para valores grandes de q, su beneficio será casi imperceptible, ya que el coste de los pasos 2 y 3) estaría dominado asintóticamente por el del paso 4). Por otro lado, hemos comprobado, a través de simulación, que el tiempo de ejecución del algoritmo tiende a empeorar empleando este método cuando q es pequeño.[65]

4) El coste asociado a este paso es $t_4(q) = c_4 rq$.

5) Empleando el procedimiento 5) en vez de 3) y 4) se obtiene un coste $t_5(q) = (c_3 + c_5)sq$ en vez de $t_3(q) + t_4(q) = (c_3 + c_4 r)q$, donde $s \leq r$ siempre, mientras $c_5 > c_4$ (porque el programa para calcular la función de masa (5) puede preparase para que realice la operación para un conjunto de vectores de frecuencias, de modo que el cálculo de $T!$ se realice una sola vez, y el programa sea llamado también una sola vez. Esto evita incurrir en ciertos costes repetidamente). La mejora en eficiencia computacional depende de la diferencia entre r y s, por tanto, depende de la muestra particular con la que se aplique Mult2.

[63] La extensión al caso impar, que es el que sucede realmente en nuestro contexto, no genera cambios en términos de coste asintótico.

[64] El algoritmo "quickshort", así como el de "búsqueda binaria", pueden consultarse en Brassard y Bratley (1997), capítulo 7. Para el "algoritmo de la casilla", véase la página 80 de esa misma referencia.

[65] Esto es habitual en muchos algoritmos para resolver distintos problemas, cuya eficiencia está demostrada analíticamente para el "caso asintótico" (cuando el tamaño del problema −que suele denotarse por n− es grande), pero no es tal para valores pequeños de n, para los que otros algoritmos, que son poco recomendables cuando crece n, producen mucho mejores resultados en tiempo de ejecución.

Obviamente, mientras q no sea un número "demasiado grande", el coste del algoritmo es perfectamente asumible, en el sentido de que el usuario no llegaría siquiera a advertir diferencia entre el tiempo de ejecución de cualquier otro contraste y el de Mult2. El problema aparece cuando q toma valores elevados. Usando la notación asintótica habitual del análisis de coste de algoritmos, se dice que el coste $z(n)$ de un algoritmo es $O(f(n))$ si $\exists n_0 \in \mathbb{N}$ y $\exists c \in \mathbb{R}$ tal que $\forall n \geq n_0$, $z(n) \leq cf(n)$. Tomando, por ejemplo, $c = c_2 + c_3 + \text{máx}(c_4 r, c_5 s)$, el coste asociado al conjunto de pasos 2) a 4) o de pasos 2) y 5) es, por tanto, $O(q)$. Incluso aunque se implementara el paso 3) con el procedimiento de búsqueda binaria explicado arriba, lo que lograría que el coste parcial asociado a dicho paso fuera $O(\log(q))$, el coste global del algoritmo (sin incluir el paso 1)) seguiría siendo $O(q)$. Para verlo, se puede utilizar la "regla del máximo" (fácilmente demostrable), según la cual si $z_1(n) \in O(f_1(n))$ y $z_2(n) \in O(f_2(n))$, entonces $z_1(n) + z_2(n) \in O(\text{máx}(f_1, f_2))$. En nuestro caso, como $\log(q) \leq q \ \forall q \in \mathbb{N}$, está claro que el coste agregado de los pasos 2) a 4) sigue siendo $O(q)$, es decir, lineal en q. Fijémonos que $q = \frac{(T+K-1)(T+K-2)...(T+1)}{K-1!}$ y, por tanto, el coste del algoritmo es aproximadamente $O(T^{K-1})$, creciendo muy rápidamente tanto con T como con K. Ésta es la conclusión que pretendíamos demostrar.

D.2. Análisis empírico

Para dar una idea práctica sobre los valores que toma en la práctica el coste del algoritmo correspondiente a Mult2, hemos calculado el tiempo de ejecución del contraste para diferentes valores de T y K, y, por tanto, de q. Se ha utilizado un computador de capacidad media, con una velocidad de procesador de 2,0 GHz y 2 GB de memoria RAM. En nuestra implementación del algoritmo, usamos el paso 5) en vez de 3)-4), procedimiento que puede ser ligeramente beneficioso cuando s es claramente menor que r, es decir, en situaciones en que la hipótesis nula es cierta (cuando la hipótesis nula es falsa, normalmente, el test tendrá que analizar el conjunto C de forma completa ($s = r$), llegando finalmente a la conclusión de que $\widehat{F}(ZZ_n^0) < \alpha$. En cambio, de ser cierta, se suele llegar a la conclusión $\widehat{F}(ZZ_n^0) > \alpha$ sin necesidad de chequear todo C ($s < r$)).

Primero, calculamos las matrices A para cada par (K, T) elegido, ejecutando el algoritmo recursivo explicado en 1), y mostramos sus tiempos de ejecución en el Cuadro 14. Recordemos que se incurriría en este tiempo una sola vez. Luego, estimamos el tiempo de ejecución del contraste (pasos 2) y 5)), conocida la matriz A, tiempo éste que es el verdadero coste computacional del test. Aplicamos el contraste sobre dos series de errores de predicción $N(0, \sigma_i^2)$ independientes. Como en cualquier algoritmo, existirán tipos de datos (en este caso, tipos de muestras) que generen costes menores que otros.[66] En este algoritmo, el coste será menor cuanto más pequeño sea sq, es decir, el número de vectores de frecuencias de C que se analicen en el paso 5). Pero la identificación del tipo de muestras que conducen a un valor menor de sq no es clara. Cuando la hipótesis nula es cierta, tendemos a no analizar todo C, mientras si es falsa, sí tendremos que hacerlo, pero el tamaño de C tenderá a ser significativamente menor, porque ZZ_n^0 tiende a ser mucho más pequeño. Normalmente, este segundo efecto es el dominante, y el "caso mejor" suele ser aquel en que la hipótesis nula es falsa (y cuanto más falsa, menor coste). Siguiendo esta argumentación, hemos distinguido dos casos: en uno de ellos, $\sigma_1^2 = \sigma_2^2 = 1$ (hipótesis nula cierta) y en el otro, $\sigma_1^2 = 3$ y $\sigma_2^2 = 1$ (hipótesis nula falsa). Esperamos que el segundo de los casos genere tiempos de ejecución algo menores, al menos para valores de q elevados, por el razonamiento anterior. Se han realizado 200 repeticiones de cada experimento, suficientes para una estimación fiable del tiempo promedio de ejecución del test. Los promedios de los tiempos obtenidos se muestran en el Cuadro 14. Se prueban los valores $K = 5$ y $K = 7$. Para el caso $K = 5$, se emplea $T = 8, 16, 24, 32, 40, 48$; en cambio, para $K = 7$, solo se probarán $T = 8, 16, 24, 32$ cuando $K = 7$, ya que el coste de cálculo de la matriz A se hace excesivamente grande si $K = 7$ y $T > 32$.

[66] Por poner un ejemplo, en un algoritmo cuya función es ordenar crecientemente el vector que recibe, el mejor caso posible podría ser aquel en el que el vector recibido está ya ordenado crecientemente, y el peor, aquel en el que el vector está justo ordenado decrecientemente. Esto ocurre en el algoritmo de ordenación por inserción.

Cuadro 14. Tiempo Promedio Ejecución Algoritmo para Test Mult2

	Cálculo Matriz A (minutos)		Resto de Pasos (segundos)				Valor de q	
			H_0 cierta	H_0 falsa	H_0 cierta	H_0 falsa		
T	$K=5$	$K=7$	$K=5$	$K=5$	$K=7$	$K=7$	$K=5$	$K=7$
8	0,001 min	0,001 min	0,016 sg	0,013 sg	0,043 sg	0,029 sg	495	3003
16	0,001 min	0,078 min	0,044 sg	0,043 sg	0,518 sg	0,512 sg	4845	74613
24	0,006 min	15,738 min	0,148 sg	0,147 sg	4,298 sg	4,264 sg	20475	593775
32	0,026 min	300,678 min	0,422 sg	0,415 sg	21,329 sg	21,505 sg	58905	2760681
40	0,088 min	-	1,001 sg	0,998 sg	-	-	135751	-
48	0,256 min	-	2,133 sg	2,108 sg	-	-	270725	-

D.3. Conclusiones

En primer lugar, el ejercicio permite comprobar que el coste del algoritmo tiende a ser mayor si la hipótesis nula es cierta que si es falsa, pero las diferencias son mínimas. En segundo lugar, se confirma que el coste de ejecución del algoritmo, en lo que concierne a los pasos 2) a 5) es aproximadamente lineal en q, tal y como habíamos demostrado en el análisis realizado en D.1. Recordemos que ésta es la parte del algoritmo que habría que ejecutar siempre. Por su parte, el coste del paso 1) (cálculo de la matriz A), en que se incurriría solo la primera vez que ejecutemos el test Mult2 para un par (K,T) dado, es una función que crece con q más rápido que linealmente. Para ver esto último, obsérvese el crecimiento del tiempo de cálculo de A entre el caso $(K=7,T=24)$ y $(K=7,T=32)$: mientras q se multiplica por 4,6, el tiempo de cálculo se multiplica por 19,1.

Pero la conclusión más relevante es respecto a los valores (K,T) para los que el tiempo de ejecución es razonable. La prueba empírica que hemos efectuado deja claro que, si $K=5$, el test Mult2 puede aplicarse para muestras de longitud menor o igual a 50 incurriendo en un coste computacional muy razonable: como máximo, se necesitarían 2 segundos para la ejecución del test una vez calculada A, aunque en la primera ejecución habría que incurrir en un coste máximo de 15 segundos para la construcción de dicha matriz. Sin embargo, si $K=7$, el coste es totalmente inaceptable para $T>32$ y poco recomendable $T\in(24,32)$. En el caso $T=24$, el tiempo de crear A alcanza los 15 minutos, y cada ejecución del contraste supone 4 segundos.

Resumiendo, debido al coste computacional, *el uso de Mult2 debe restringirse a* $K=5$, $T\leq 50$ y $K=7$, $T\leq 25$, si la muestra no presenta autocorrelación.[67] Si existe autocorrelación de orden r, las longitudes muestrales máximas para aplicar el test serían $r+1$ veces las mencionadas, ya que la muestra original de tamaño T se desglosa en $r+1$ submuestras de tamaño $T/(r+1)$, al aplicar el procedimiento de contrastes múltiples bajo la cota de Bonferroni.

[67] Por supuesto, para $K<5$, el test se puede aplicar con muestras de cualquier longitud T sin ningún problema de coste computacional.

E. Apéndice: Características de la aleatorización aplicada en los tests

Para la obtención de las estimaciones de tamaño y potencia (SAP) del test Mult2 y de las versiones exactas de los tests de Signos y Wilcoxon se ha aplicado el mecanismo de aleatorización, cuya descripción en el caso general fue presentada en el Apéndice A del Capítulo 1 de esta Tesis. En el segundo punto del apartado 3.2.5 del capítulo actual se expusieron brevemente cuatro cuestiones relevantes para la implementación e interpretación del método de la aleatorización en las simulaciones de este documento. La ampliación de dichas cuestiones es el objeto de la primera parte de este Apéndice. Por otro lado, el cuarto de aquellos comentarios argumentaba la gran relevancia que tiene la probabilidad de aleatorización P_d en cada test, de cara a evaluar su verdadera aplicabilidad práctica y el "realismo" de sus propiedades teóricas. Pues bien, en la segunda parte de este Apéndice ofrecemos las estimaciones de dicha probabilidad, obtenidas a través de las simulaciones correspondientes a los ejercicios de tamaño de la sección 4. Dichas estimaciones son fundamentales para mostrar la preferencia en la práctica por el test Mult2, respecto a los otros dos tests discretos evaluados, Signos y Wilcoxon, tal y como se afirmó en el punto 2 del apartado 4.2. Además de presentar las estimaciones, en la segunda parte del Apéndice explicamos los detalles sobre su cálculo.

E.1. Cuestiones relacionadas con la implementación del procedimiento

Desarrollamos y ampliamos aquí las cuatro cuestiones relativas a la aplicación del procedimiento de aleatorización en nuestro contexto que habían sido presentadas con brevedad en el apartado 3.2.5:

a) La primera cuestión trataba sobre la posibilidad de utilizar en las simulaciones niveles de significación "peculiares" para los tests discretos (correspondientes a puntos de discontinuidad de la distribución de contraste), evitando así la necesidad de aleatorizar estos tests. Sin embargo, descartábamos tal opción, y las razones son las siguientes:

− Primero, porque deseamos comprobar el funcionamiento de los contrastes Mult2, Signos y Wilcoxon bajo niveles de significación habituales en la práctica, y no bajo niveles que no van a ser utilizados por los usuarios.

− Segundo, porque la utilización de dichos niveles de significación para estos tres contrastes, impediría una comparación directa con las propiedades de los otros tests, que se evalúan bajo el nivel de confianza convencional $\alpha = 0{,}10$.

− Finalmente, porque dichos niveles son particulares de cada T en el caso de los tests Signos y Wilcoxon, y de cada vector $(p', b', T)'$ en el caso de Mult2, ya que la distribución F_{ZZ_n} de dicho test depende de esos tres parámetros. De este modo, habría que situar un nivel para cada uno de estos casos, haciendo que las tablas de resultados resultaran innecesariamente incómodas de interpretar.

b) La segunda se refería a la discrepancia entre las distribuciones exactas correspondientes a los contrastes Wilcoxon y Mult2 y las usadas en su implementación, discrepancia que motiva la posibilidad de que se detecte sesgo en el tamaño de estos tests, pese a haber sido aleatorizados. El caso de Mult2 ya fue explicado en el apartado 3.2.5. Por su parte, la razón por la que la distribución teórica del estadístico de Wilcoxon puede no coincidir con la verdadera es que la derivación de la distribución para muestras finitas del estadístico de contraste S_3 empleaba el supuesto de continuidad en cierta variable (véase (3)), y esa condición no se cumple en nuestro contexto. Este asunto fue adelantado en el punto 4) del apartado 3.1.2, donde se presentaba dicho test.

c) En tercer lugar, se comentó que la relevancia del problema que la aleatorización trata de resolver −la imposibilidad de construir regiones críticas de tamaño α− disminuía al aumentar el número de elementos del soporte de la variable aleatoria discreta usada en el test. La causa obvia es que un aumento en dicho número implica un incremento en el número de "escalones" de la función de distribución y, al ser ésta no decreciente y estar acotado el rango de sus posibles imágenes dentro del intervalo $[0, 1]$, su altura debe disminuir. Respecto a los tests de Signos y Wilcoxon, ya se comentó en 3.2.5 que la cardinalidad del soporte de sus estadísticos de contraste depende únicamente de T. Por su parte, el soporte del estadístico de contraste de Mult2 es más difícil de determinar, pero su cardinalidad depende positivamente de T y del número de pérdidas definido K, aunque también está relacionado con el valor numérico del vector b.

d) Finalmente, se definió una propiedad fundamental en los tests discretos: que la probabilidad $P_d = P(\lambda = \lambda_d)$ no fuera elevada. Se comentó que pensábamos que P_d probablemente sería menor en Mult2 que en Signos y Wilcoxon. El motivo es que, por lo expuesto en el punto c), P_d tiende a ser menor cuanto mayor sea la cardinalidad del soporte del estadístico de contraste, ya que, así, también será menor la masa de probabilidad media de los puntos del soporte. Por la construcción de los tests, para la misma longitud muestral T, el soporte de Mult2 tiende a ser bastante más "denso" que el de Wilcoxon, y éste, a su vez, más que el de Signos.

Por otro lado, existen dos preguntas relevantes respecto a la estimación de P_d en los tests discretos que nos ocupan, a saber:

– ¿Por qué hay que estimar P_d, si solo es la probabilidad asociada a un punto de la distribución de contraste, supuestamente conocida? En el caso de Mult2, la respuesta es obvia, hay que estimarlas porque la distribución de contraste F_{ZZ_n} no se conoce, y precisamente el test consiste esencialmente en estimar dicha distribución (una parte de ella, en realidad). Sin embargo, en Signos y Wilcoxon conocemos la distribución teórica del contraste. Enseguida aclararemos porqué se requiere una estimación.

– ¿Cuántas P_d hay que estimar en los ejercicios de simulación? Obviamente, las probabilidades P_d dependen de las mismas variables que caracterizan la distribución de contraste, y, por tanto, hay que estimar una probabilidad P_d por cada distribución de probabilidad distinta que deba usarse en los experimentos, correspondiente al test en cuestión. Bien, en el caso de Mult2, ya se comentó en el primer punto del apartado 3.2.2 que la distribución F_{ZZ_n} depende de todos los componentes del vector (p', b', T). En nuestras simulaciones, b permanece constante, mientras la longitud muestral T varía, y p es diferente para cada escenario bajo el que se simulan los errores de previsión, es decir, varía con θ y ρ. Por tanto, está claro que, para el test Mult2, hay que estimar P_d para cada valor de la terna (θ, ρ, T), es decir, para cada uno de los diseños del experimento. En cambio, la distribución exacta del test de Signos y la del test de Wilcoxon están unívocamente caracterizadas por T. Sin embargo, por la razón que veremos a continuación, la probabilidad P_d varía también con θ y ρ, igual que en Mult2.

La razón a la que venimos refiriéndonos tiene que ver con la implementación de Signos y Wilcoxon, que eliminan de la muestra original aquellas observaciones $d_t = 0$, calculándose su estadístico de contraste sobre la muestra modificada de longitud T', y, por supuesto, utilizando la distribución teórica asociada a T', no a T. Entonces, en cada simulación correspondiente a un diseño de longitud muestral teórica T, no se conoce a priori el número de observaciones $d_t = 0$ que se eliminarán, y, por tanto, no se sabe a priori la función de distribución a utilizar. En consecuencia, no queda más remedio que estimar P_d a través de las simulaciones. El valor teórico de P_d en Signos o Wilcoxon, para un T dado, será la media ponderada de las probabilidades de aleatorización de las distribuciones correspondientes a todas las longitudes muestrales $t \leq T$, probabilidades que denotaremos por $P_{d,t}$. Por su parte, la ponderación es, a su vez, la probabilidad de que la muestra inicial de T observaciones quede en solo t observaciones, tras eliminar aquellas de valor cero. Denotamos estas probabilidades por $\eta_{T,t}$. Es decir, $P_d = \eta_{T,1}P_{d,1} + \eta_{T,2}P_{d,2} + ... + \eta_{T,T}P_{d,T}$. En esta situación, está claro que P_d debe estimarse, puesto que no conocemos a priori las probabilidades $\eta_{T,t}$, ni tampoco está garantizado conocer las verdaderas $P_{d,t}$.[68] Por otro lado, las $\eta_{T,t}$ varían según sea el escenario de simulación (ρ, θ), ya que éstos definen p, por lo que habrá que estimar un P_d para cada terna (θ, ρ, T), es decir, para cada diseño del experimento.

E.2. Estimaciones de la probabilidad de aleatorización

En el caso de Mult2, la estimación de la probabilidad P_d asociada al diseño (θ, ρ, T) será $\widehat{P}_d = \widehat{F}_{(d)} - \widehat{F}_{(b)}$, donde $\widehat{F}$ denota la estimación de la función de distribución F_{ZZ_n} del test, correspondiente al escenario de simulación (θ, ρ) y longitud muestral T. Dicha estimación no será sino la función de distribución empírica del estadístico de contraste, obtenida a partir de valores observados de éste en las 10000 realizaciones del ejercicio.

En cambio, para los tests Signos y Wilcoxon, la probabilidad P_d asociada a un diseño (θ, ρ, T) se estimará siguiendo la expresión $P_d = \eta_{T,1}P_{d,1} + \eta_{T,2}P_{d,2} + ... + \eta_{T,T}P_{d,T}$, donde las estimaciones de $P_{d,t}$ y $\eta_{T,t}$ serán:

[68] Por las razones expuestas en el punto b) del apartado E.1, la verdadera distribución del contraste puede diferir de la teórica, al menos en el test de Wilcoxon. Por eso, conviene estimar $P_{d,t}$, en vez de tomarla de la distribución teórica correspondiente, al menos si se dispone de suficientes observaciones para hacerlo con precisión.

a) $\widehat{P}_{d,t} = \widehat{F}^t_{(d)} - \widehat{F}^t_{(b)}$, donde $\widehat{F}^t$ denota la estimación de la función de distribución del test para longitud muestral t, asociada al escenario de simulación (θ, ρ).

b) $\widehat{\eta}_{T,t} = \frac{N_t}{N}$, siendo N el número total de repeticiones del ejercicio (en nuestro caso, siempre 10000), mientras N_t denota el número de repeticiones del ejercicio en las que la muestra original, de longitud T, finalmente se transformó en una muestra de longitud t, tras eliminar las observaciones nulas.

En este caso, la función de distribución empírica $\widehat{F}^t$ se habrá estimado en base a los N_t valores observados del estadístico de contraste correspondientes a muestras de longitud inicial T pero longitud final t. Cuando $N_t < 1000$, utilizamos la distribución teórica del contraste, para longitud t, para evitar estimaciones poco precisas.

Los siguientes comentarios adicionales respecto a la estimación de P_d son relevantes:

1. En el caso del test de Wilcoxon, se usó la versión asintótica siempre que $T > 20$, lo que significa que no se empleó aleatorización en estas situaciones. Por ello, la probabilidad de aleatorización estimada para Wilcoxon disminuye mucho cuando T es elevado, ya que se está utilizando $\widehat{P}_{d,t} = 0$ para cualquier $t > 20$. Es decir, que el buen resultado aparentemente alcanzado por el test en dichas longitudes muestrales en cuanto a probabilidad de aleatorización no es producto de ninguna propiedad real del test, sino de la implementación que hemos realizado.

2. La estimación de P_d podría haberse llevado a cabo, para cada uno de estos tres tests, simplemente contabilizando el porcentaje de repeticiones de cada ejercicio en que se requirió un experimento de Bernouilli para resolver el contraste. Hemos llevado a cabo también dicha contabilización, y los resultados son muy similares a los presentados en la Tabla 19, bajo el método descrito arriba.

3. Se adjuntan solamente las estimaciones $\widehat{P}_d$ que corresponden a los ejercicios de tamaño. Para los ejercicios SAP, la estimación por el método utilizado se complicaría, ya que la probabilidad de aleatorización está asociada a la distribución bajo la hipótesis nula, mientras los estadísticos muestrales almacenados son generados por la verdadera distribución (distinta de la de la hipótesis nula), y conducirían a un cálculo incorrecto.

Tabla 19. Estimaciones Frecuencia (%) Randomización.
Ejercicios de Tamaño

		$\theta = 0$ $(h = 1)$			$\theta = 1$ $(h = 2)$		
	T	Signos	Wilcoxon	Mult2	Signos	Wilcoxon	Mult2
$\rho = 0,0$	8	27,4	20,5	6,5	29,0	21,2	6,4
	16	16,8	3,4	4,9	23,0	14,9	6,2
	24	13,5	2,1	3,9	24,8	13,3	6,6
	32	11,2	1,3	3,2	20,3	5,5	6,0
	40	9,5	0,7	3,0	17,6	3,1	5,4
	48	8,9	0,2	3,1	15,7	2,3	4,7
$\rho = 0,5$	8	35,9	32,8	9,9	37,1	33,6	5,2
	16	19,9	7,4	6,3	30,9	27,9	7,7
	24	14,9	2,5	4,3	31,6	25,5	7,2
	32	12,9	1,7	3,8	24,3	12,4	5,9
	40	11,5	1,3	3,5	20,6	6,2	6,1
	48	10,1	0,8	3,0	18,1	3,5	5,4
$\rho = 0,9$	8	53,3	53,3	8,5			
	16	39,9	37,0	8,7			
	24	28,2	20,5	6,9			
	32	21,8	10,6	5,6			
	40	18,3	5,9	5,0			
	48	16,2	3,7	4,5			

Los resultados están asociados a las Tablas 5 y 6.

Referencias

[1] Ash, J.C.K., Smyth, D.J. y Heravi, S.M. (1998). Are OECD Forecasts Rational and Useful?: a Directional Analysis, *International Journal of Forecasting* 14, 381-391.

[2] Birchenall, C.R., Jessen, H. y Osborn, D.R. (1996). Predicting US Business Cycle Regimes, *Journal of Business and Economic Statistics*, 17, 313-323.

[3] Brassard, G. y Bratley, P. (1997). Fundamentos de Algoritmia. Prentice Hall (Madrid).

[4] Campbell, B. y Ghysels, E. (1995). Is the Outcome of the Federal Budget Process Unbiased and Efficient? A Nonparametric Assesment, *Review of Economics and Statistics* 77, 17-31.

[5] Chinn, M. y Meese, R.A. (1991). Banking on Currency Forecasts: Is Change in Money Predictable?, *Journal of International Economics* 38, 161-178.

[6] Clements, M.P. y Hendry, D.F. (1993). On the Limitations of Comparing Mean Square Forecast Errors, *Journal of Forecasting* 12, 617-637.

[7] Dell'Aquila, R. y Ronchetti, E. (2004). Robust Tests of Predictive Accuracy, *Metron - International Journal of Statistics* 62, 161-184..

[8] Diebold, F.X. y Mariano, R.S. (1995). Comparing Predictive Accuracy, *Journal of Business and Economic Statistics*, 13, 253-263.

[9] Greer, M. (2003). Directional Accuracy Tests of Long-Term Interest Rate Forecasts, *International Journal of Forecasting* 19, 291-298.

[10] Harvey, D.I., Leybourne, S.J. y Newbold, P. (1997). Testing the Equality of Prediction Mean Squared Errors, *International Journal of Forecasting*, 13, 281-291.

[11] Leicht, G. y Tanner, J.E. (1991). Econometric Forecast Evaluation: Profits versus the Conventional Error Measures, *American Economic Review* 81, 580-590.

[12] Leitch, G. y Tanner, J.E. (1995). Professional Economic Forecasts: Are they Worth their Costs?, *Journal of Forecasting*, 14, 143-157.

[13] Schnader, M.H. y Stekler, H.O. (1990). Evaluating Predictions of Change, *The Journal of Business* 63, 1, 99-107.

[14] Stekler, H.O. (1994). Are Economic Forecasts Valuable?, *Journal of Forecasting* 13, 495-505.

[15] West, K.D., Edison, H.J. y Cho, D. (1993). A Utility-Based Comparison of Some Models of Exchange Rate Volatility, *Journal of International Economics* 35, 23-46.

CAPÍTULO 3

EFECTO DE LA INCERTIDUMBRE PARAMÉTRICA SOBRE CONTRASTES PARA LA EVALUACIÓN O COMPARACIÓN DE CAPACIDAD PREDICTIVA, BAJO FUNCIÓN DE PÉRDIDA DISCRETA

1. Introducción

En el primer y segundo capítulo de esta tesis se han ido presentando algunos tests que utilizan funciones de pérdidas para evaluar la capacidad predictiva de un modelo (Capítulo 1) y para comparar dicha capacidad entre dos modelos (Capítulo 2). Todos los contrastes estaban derivados bajo el supuesto de que los modelos de predicción utilizados no han sido estimados, es decir, asumiendo que se conocen los verdaderos valores (β^*) del vector paramétrico β que aparezca en su especificación. Sin embargo, cuando tal hipótesis no se cumple, como sucede casi en la totalidad de las aplicaciones prácticas, los estadísticos involucrados en los tests están afectados por dos fuentes de incertidumbre. La primera está presente en todos los casos, aunque se conozca β^*, y se debe simplemente a la condición de variable aleatoria que tienen los datos para los que se pretende hacer previsión. Esta incertidumbre, que se conoce como error en el muestreo, es la única que se está teniendo en cuenta en la construcción de los tests en los capítulos anteriores. La segunda fuente de incertidumbre, hasta ahora no considerada, procede de la estimación paramétrica. La denotaremos por IP ("incertidumbre paramétrica"). La aplicación de estos contrastes de forma directa, es decir, sin incluir correcciones motivadas por la existencia de una estimación previa a la previsión, puede llevar a distorsiones importantes en sus propiedades estadísticas de tamaño y potencia.

West (1996) es la referencia básica para este tipo de cuestiones. Presenta un teorema que especifica la expresión de la matriz de varianzas y covarianzas asintótica correcta bajo IP − matriz que se denotará por Ω − para una amplia familia de tests que utilicen una función de pérdidas f, y que contrasten la esperanza matemática de f a través del estadístico media muestral $\overline{f}$. En esta familia se incluyen el test de Signos, P-T (Pesaran-Timmermann) y nuestro contraste C1 en cualquiera de sus dos versiones − respecto a los tests para evaluar la habilidad predictiva de un modelo (Capítulo 1) −, y los tests de Signos, Wilcoxon, DM (Diebold y Mariano (1995)) y nuestro contraste Mult2-aprx[1] − respecto a los tests de comparación de capacidad predictiva entre dos modelos (Capítulo 2) −. El teorema de West (1996) es válido bajo ciertas condiciones, entre las que destacan (a) diferenciabilidad de f respecto a β, (b) estacionariedad de f y de cierta función de los regresores y (c) una condición analítica sobre la expresión del estimador de β. La condición (a) restringe el conjunto de funciones de pérdida, el supuesto (b) restringe el tipo de modelo de previsión y el supuesto (c) limita la clase de estimadores a los que dicho teorema aplica. La matriz de covarianzas Ω es de la forma $\Omega = S_{ff} + FQ$, siendo S_{ff} la matriz de covarianzas derivada bajo ausencia de IP (β^* conocido) y $F = E\left[\frac{\partial f(\beta)}{\partial \beta}\right]_{\beta=\beta^*}$, mientras la matriz Q está relacionada, entre otras cosas, con la matriz de varianzas y covarianzas del estimador de β.

El resultado de West (1996) ha sido extendido posteriormente en varias direcciones, para cubrir situaciones de predicción en las que el teorema de West (1996) no aplica. Por un lado, el efecto de la estimación paramétrica en presencia de variables exactamente integradas, casi integradas y cointegradas (incumpliendo, por tanto, la condición (b) anterior) es tratado en Berkowitz y Giorgianni (2001), Rossi (2005) y Corradi, Swanson y Olivetti (2001). No obstante, esta extensión no será tratada en nuestro trabajo, que se circunscribe al caso de variables estacionarias. Por otro lado, y éste es un punto clave para la investigación que presentamos, McCracken (2000) generaliza el teorema de West al caso de funciones de pérdida que no son diferenciables en β (incumpliendo el supuesto (a) de aquel teorema), siempre que la esperanza matemática de f sí lo sea. La expresión que obtiene para Ω es la misma que en West (1996), solo que ahora $F = \left[\frac{\partial E f(\beta)}{\partial \beta}\right]_{\beta=\beta^*}$.

[1] Mult2-aprx es un test que presentamos en el capítulo 2 (versión asintótica de Mult2), cuya única diferencia respecto al test DM aplicado con función de pérdidas discreta es el cálculo de la varianza del estadístico.

El problema de la aplicación práctica de los resultados de West (1996) y McCracken (2000) es la complejidad de estimar la parte de la varianza de los contrastes que se debe a la estimación paramétrica del modelo, FQ. Por un lado, por razones que se comprenderán cuando se presente la expresión de la matriz Q, la estimación de ésta es, como mínimo, muy laboriosa, e incluso podría ser imposible si el usuario del test no es quien generó las previsiones y, por tanto, desconoce los detalles de la estimación. Algo similar puede decirse de la estimación de F cuando la función f es diferenciable. Pero el verdadero problema emerge cuando f no es diferenciable, porque entonces la derivación de la expresión analítica de F suele ser muy compleja, y habitualmente, requiere hacer supuestos sobre la distribución de las variables explicativas del modelo de previsión. Véase McCracken (2004) para profundizar sobre este asunto.

Por todo esto, una situación muy deseable cuando se predice a través de modelos estimados es que la IP asociada a la estimación de los modelos sea irrelevante en términos de la distribución asintótica de los contrastes, es decir, que la matriz Ω coincida con S_{ff} (matriz ésta que no ofrece problemas de estimación). Nos referiremos a esta circunstancia como "irrelevancia asintótica de la incertidumbre paramétrica", y la denotaremos por IAIP. Existen tres maneras por las que pudiera cumplirse la condición $FQ = 0$, pero dos de ellas aparecen con muy poca frecuencia en la práctica (véase West (1996), pág. 1073). La tercera posibilidad es la verdaderamente trascendente en este análisis, y consiste en que la matriz F sea nula. Desgraciadamente, son pocos los casos identificados en la literatura en los que se verifica la condición $F = 0$, a saber:

i) Si la función que se optimiza (intramuestralmente) para obtener un estimador consistente de β es la misma que la función de pérdidas f empleada para la valoración (extramuestral) de las previsiones.

ii) Si f es el cuadrado del error de previsión (denotaremos dicha función por SE) y la media de los errores de previsión condicionada en los valores del mismo periodo de las variables explicativas del modelo de previsión es nula. En modelos lineales, una condición igualmente válida es que los errores y las variables explicativas estén incorreladas contemporáneamente.

En la práctica, es difícil pensar en casos en que la condición i) se verifique, salvo cuando ambas funciones son la cuadrática, es decir, la pérdida es SE y el estimador es OLS,[2] siempre que éste sea consistente. En consecuencia, estas condiciones podrían prácticamente sintetizarse en ii).[3]

iii) Una tercera posibilidad es que f sea el valor absoluto del error de previsión (denotaremos dicha función por AE), exista incorrelación entre errores de previsión y regresores y, además, la distribución de los primeros sea simétrica. Esta condición fue mostrada en McCracken (2000).[4]

Por tanto, incluso aunque no hubiera problemas de correlación entre errores y regresores y el supuesto de simetría en la distribución de los errores fuera asumible, es probable que haya que estimar FQ si se usan funciones de pérdida que no sean las estándar, SE o AE, con el elevado coste que conlleva. Además, si f es no diferenciable, habrá que añadir a esto las notables dificultades ya citadas de calcular la expresión de la matriz F.

Teniendo en cuenta lo comentado sobre la complejidad de estimar Ω, y sobre el escaso conjunto de situaciones que garantizan IAIP, es natural preguntarse si utilizar S_{ff} como matriz de covarianzas de los contrastes en un contexto en el que no se cumple IAIP es un procedimiento con efectos graves sobre las propiedades de los tests. Para comprobarlo, West (1996) diseña un experimento de simulación en el que existen dos modelos de predicción con la misma capacidad predictiva, pero cuyos errores están correlados con los regresores del propio modelo. Los modelos son comparados a través del test DM, usando como función de pérdida el SE, y estimando la varianza del test por S_{ff}, en vez de por la varianza correcta Ω. No se está verificando la condición ii), y, por consiguiente, se trata de una situación en la que la IP será relevante asintóticamente. Pues bien, el sesgo en el tamaño de DM que obtiene West (1996) en este experimento sobrepasa el 40 % en algunos tamaños muestrales , cuando el tamaño teórico fijado era de 5 %. Nosotros hemos comprobado que el sesgo que se obtendría en este mismo ejercicio usando AE sería de cuantía similar al obtenido bajo SE. Las cifras anteriores pueden dar una idea de la importancia del asunto al que dedicamos este trabajo.

[2] O cualquier método que minimice la suma residual, como NLLS (en modelos no lineales), o MV bajo normalidad en las perturbaciones.

[3] Porque para que un estimador OLS sea consistente debe verificarse precisamente que sea nula la media de los errores condicionada a los regresores.

[4] Las notaciones SE y AE se refieren a "Square Error" y "Absolute Error".

Para completar una panorámica general sobre el efecto de la IP en la clase de tests a los que se refiere toda esta literatura, falta un último punto, de gran importancia. Esencialmente, los contrastes en cuestión pueden ser contrastes para evaluar la capacidad predictiva de un solo modelo, o para comparar la de dos o más. Sin embargo, hay un caso en el que los resultados citados hasta ahora no aplican: si se trata de tests de comparación predictiva e involucran modelos anidados. En tal caso, no está garantizada siquiera la convergencia de los estadísticos de contraste a una distribución, o, en caso de converger, la distribución resultante podría ser degenerada o, como mínimo, no normal. En el contexto de modelos anidados, no existe un resultado en la literatura tan general como los de West (1996) o McCracken (2000). Sin embargo, sí han ido apareciendo trabajos que tratan la incorporación de la IP cuando se comparan exactamente dos modelos anidados, pero solo desarrollan teoría para contrastes concretos. En este sentido, probablemente, la referencia fundamental es McCracken (2007). Éste centra su análisis en el caso del test DM y función de pérdida diferenciable e igual a la empleada en la estimación. Para dicho contexto (y bajo algún supuesto adicional), demuestra la convergencia en distribución del estadístico y deriva la distribución asintótica, que, desde luego, no es estándar, proporcionando el autor los percentiles necesarios para la implementación del test. Clark y McCracken (2001) y Clark y West (2007) aportan también resultados de interés en el marco de modelos anidados, para otros tests particulares, diferentes de DM.

La motivación de nuestro trabajo es comprobar el efecto que tiene la IP sobre el tamaño de tests de los capítulos anteriores, cuando se emplea una función de pérdidas discreta (el dominio de datos y previsiones $R \times R$ se discretiza en n^2 cuadrantes, asignando una pérdida a cada uno), que es un caso concreto de función "continua a trozos", y, por tanto, no diferenciable. Restringiremos el análisis al caso de modelos lineales estacionarios y, dividiremos en dos posibles ámbitos predictivos: (1) Modelos no anidados, (2) Modelos anidados.

En el caso (1), el resultado de McCracken (2000) aplica para nuestra función de pérdidas. El trabajo que nos proponemos es el siguiente:

a) Obtener la expresión analítica de la matriz F bajo f discreta. Recordemos la dificultad inherente a este objetivo cuando f es no diferenciable.

b) Una vez logrado el paso anterior, trataremos de comprobar si existen situaciones en las que, bajo nuestra pérdida discreta, se verifica la condición $F = 0$ (que garantiza IAIP), y, en caso afirmativo, las caracterizaremos. De nuevo, recuérdese la escasez de casos identificados en la literatura para los que aplique tal propiedad.

c) Finalmente, analizaremos los resultados de los tests en situaciones diferentes a las encontradas en b), es decir, en las que la IP sea relevante, pero bajo nuestra función de pérdidas discreta. Igual que se hizo en West (1996) para el SE, el propósito es cuantificar la distorsión que se produce en las propiedades de los contrastes si se emplea S_{ff} en vez de Ω, pero cuando la pérdida es discreta, en vez del SE ó el AE. Para ello, realizaremos los ejercicios de simulación que procedan, centrándonos en el contraste DM, empleado, alternativamente, bajo las tres funciones de pérdida recién citadas.

En el caso (2), no existe literatura que defina la expresión analítica de Ω para funciones de pérdida no diferenciables. Por ello, los pasos a) y b) no tienen sentido ahora, pero sí el objetivo c). Así que nos conformaremos con realizar experimentos de simulación usando S_{ff} en la aplicación del contraste DM, para orientar sobre la gravedad de no corregir los tests por la IP cuando se comparan modelos estimados que estaban anidados.

Los resultados que obtenemos son muy interesantes. En el contexto (1), los tres pasos anteriores se resuelven exitosamente. Primero, logramos derivar la expresión analítica de F, punto que, por sí mismo, no es desdeñable, tal y como hemos venido explicando en esta Introducción. No obstante, la expresión encontrada es poco manejable y, como se apuntaba en McCracken (2004), requiere supuestos sobre la distribución de probabilidad de las variables involucradas. En segundo lugar, deducimos que, bajo ciertas condiciones razonables sobre la especificación concreta de la función discreta f, se verifica la condición $F = 0$ con tal de que haya independencia entre regresores y errores de previsión y la distribución de éstos últimos sea simétrica. Así que hemos logrado un resultado análogo a iii), para pérdidas discretas, aunque necesitamos exigir independencia entre las variables citadas, en vez de solo incorrelación. En tercer lugar, realizando el mismo experimento de simulación de West para modelos con variables explicativas correlacionadas contemporáneamente con los errores, y aplicando la expresión obtenida en a) para F, obtenemos el sesgo en tamaño, tanto en muestras finitas como asintóticamente, que se produce en los tests de interés si se ignora la IP (es decir, si se emplea S_{ff}) bajo f discreta. Éste resulta ser considerablemente inferior al que se tiene bajo SE o AE (para un tamaño teórico del 5 %, el tamaño asintótico obtenido en

un ejercicio representativo está en torno al 7,6 %, frente al 28 % y 24 % que se producen para SE y AE, respectivamente). En el contexto (2), los ejercicios de simulación muestran que no corregir el test DM por la IP asociada a la estimación del modelo tiende a generar un sesgo inaceptable en el tamaño si se empleó la función SE, mientras, por el contrario, las propiedades del test se ven bastante menos afectadas si las pérdidas son de tipo discreto. Es más, a diferencia de lo que ocurre bajo pérdida cuadrática,[5] el estadístico DM conserva aproximadamente la normalidad asintótica en muchos de los diseños predictivos definidos, si f es discreta.

La conclusión general es que, si se usa una función de pérdidas discreta, normalmente, el analista podría aplicar tests como DM sin preocuparse de modificar la varianza habitual utilizada en su estadístico ni de utilizar una distribución de contraste no estándar, y las propiedades del test no se alterarán excesivamente. Es decir, nuestros resultados apuntan a que la función discreta suele garantizar que el efecto de la estimación paramétrica sobre los tests en cuestión sea leve o nulo en la mayor parte de situaciones predictivas con variables estacionarias, algo que no ocurre ni con el SE. Dada la complejidad y coste de estimar la matriz Ω, pensamos que el resultado tiene un valor considerable.

El capítulo se organiza como sigue: en la sección 2 se presenta una breve revisión de la literatura relevante al caso que nos ocupa, cuya notación se usará a lo largo del documento, y, además, se incluye un repaso de la definición de la función de pérdida discreta, ya introducida en los capítulos 1 y 2 de esta Tesis. En las secciones 3 y 4 se analizan las dos categorías de modelos de previsión en que se dividió el análisis, obteniendo los resultados respecto al efecto que se produce en las propiedades de los tests al ignorar la IP, en cada caso. La parte correspondiente a modelos lineales no anidados (sección 3), es la que exponemos con mayor grado de detalle. Presentamos teoremas que caracterizan algunas situaciones que conducen a IAIP para las funciones de pérdidas discretas. Por su parte, para otros casos en los que dichas situaciones no se cumplen, ofrecemos estudios de simulación, tanto para muestras finitas como para el caso asintótico. Finalmente, en la sección 5 se resumen las conclusiones más relevantes del capítulo.

[5] A veces nos referiremos a la pérdida SE por el término "pérdida cuadrática".

2. Contexto general. Revisión de la literatura básica

Sea M_i un modelo de previsión $y_{t+h} = \varphi_i(\beta_i^*, X_{it}) + u_{it+h}$ que genera previsiones $v_{i,t+h}$ en t para la variable y_{t+h}, siendo β_i el vector paramétrico que aparece en la especificación del modelo y β_i^* su valor verdadero, y siendo X_{it} un vector de variables observables/conocidas en t. Las variables incluidas en X pueden ser tanto regresores exógenos Z como retardos de y; además, el subíndice temporal t solo indica que los datos de X son conocidos en t, pero podrían corresponder a periodos posteriores, incluido $t+h$. Por su parte, u_{it+h} simboliza el error de previsión asociado al modelo si β_i^* fuera conocido. Es relevante entender que u_{it+h} incluye la perturbación del modelo correcto y el error derivado de una especificación incorrecta de M_i. Por ejemplo, supongamos que la variable y_{t+h} responde realmente al modelo $y_{t+h} = z_{1t} + 2z_{2t} + \varepsilon_t$, pero el modelo de previsión especificado es $y_{t+h} = \beta_0 + \beta_1 z_{1t} + u_t$. En tal caso, el error de previsión será $u_t(\beta^*) = 2z_{2t} + \varepsilon_t$.

Supondremos que se dispone de una muestra de tamaño $T+h$ para y, X. La previsión $v_{i,t+h}$ realizada en t para y_{t+h} se construye como $v_{i,t+h} = \varphi_i(\widehat{\beta}_{it}, X_{it})$, donde $\widehat{\beta}_{it}$ es la estimación de β_i en base a los datos muestrales correspondientes a los periodos $s = 1, ..., t$. El mínimo de observaciones para estimar el modelo es R y cada periodo se añade una observación más para la estimación, hasta el periodo T, el último en el que se hace previsión (y_{T+1}, ..., y_{T+h} solo se emplean para la comparación correspondiente con sus predicciones asociadas). El número total de predicciones será P, verificándose $R + P - 1 = T$. Es decir, en $t = R$ se realiza la primera previsión, correspondiente a y_{R+h}, estimando β_i con la muestra hasta $t = R$; en $t = R + 1$, la segunda, estimando con los datos hasta $t = R + 1$; así sucesivamente, hasta realizar la P-ésima previsión, correspondiente a $y_{R+P+h-1}$ (último dato disponible), con los datos hasta $t = T = R + P - 1$.[6]

Considérense l modelos de predicción. Se agruparán todos los vectores β_i en un solo vector $\beta = (\beta_1', \beta_2', ..., \beta_l')'$ de dimensión $k \times 1$, cuyo verdadero valor se denotará por β^*. Se procede de manera análoga con las estimaciones $\widehat{\beta}_{it}$ y los regresores X_{it}, compactándose en $\widehat{\beta}_t$ y X_t, respectivamente. Se define una función vectorial $l \times 1$ $f_{t+h}(y_{t+h}, \beta^*, X_t)$, cuya componente i-ésima es el valor de la pérdida en $t+h$ asociada al modelo de previsión M_i.[7] La estimación paramétrica de $f_{t+h}(y_{t+h}, \beta^*, X_t)$ será $f_{t+h}(y_{t+h}, \widehat{\beta}_t, X_t)$. Para simplificar notación, $f_{t+h}(y_{t+h}, \beta^*, X_t)$ se simbolizará por $f_{t+h}(\beta^*)$. Por ejemplo, sean dos modelos de previsión uniecuacionales y lineales de la forma $y_t = Z_{it}'\beta_i^* + u_{it}$, con $i = 1, 2$, y sea la función de pérdida SE, entonces $f_{t+1}(\beta^*) = \left((y_{t+1} - Z_{1t+1}'\beta_1^*)^2, (y_{t+1} - Z_{2t+1}'\beta_2^*)^2\right)' = \left(u_{1t+1}^2, u_{2t+1}^2\right)'$ y $f_{t+1}(\widehat{\beta}_t) = \left((y_{t+1} - Z_{1t+1}'\widehat{\beta}_{1t})^2, (y_{t+1} - Z_{2t+1}'\widehat{\beta}_{2t})^2\right)'$.

Pues bien, el tipo de tests estadísticos por los que estamos interesados son aquellos que contrastan el valor de $E(f_t)$ y requieren únicamente la distribución asintótica del estimador $\overline{f} - E(f_t)$, siendo $\overline{f} = P^{-1} \sum_{t=R}^{T} f_{t+h}(\widehat{\beta}_t)$. Véase que DM es un ejemplo de este tipo de tests.

Queremos aclarar que, aunque en los capítulos 1 y 2 de esta Tesis Doctoral se ha utilizado la notación g para referirnos a la función de pérdidas, en este capítulo usamos la letra f. Precisamente ésta era la letra utilizada para designar las funciones de comparación de pérdidas en el Capítulo 2 (por ejemplo, diferenciales $f = g_2 - g_1$), por lo que puede haber cierta tendencia al equívoco. El motivo de emplear ahora f para designar las funciones de pérdida es simplemente hacer nuestra notación del capítulo actual consistente con la de toda la literatura que nos ocupa en este capítulo, en la que se utiliza dicha letra sistemáticamente.

2.1. Literatura básica

A continuación presentamos aquellos resultados fundamentales de la literatura sobre incorporación de IP en la familia de contrastes recién descrita, que estén directamente relacionados con nuestro trabajo. En la Introducción ya se mencionaron muchos de estos resultados, pero sin entrar en expresiones analíticas, y, en cambio, haciendo hincapié en sus implicaciones, al contrario de lo que haremos ahora. Por ello, recomendamos al lector revisar los comentarios efectuados en la Introducción, a la vez que se avance en

[6] Éste es el llamado método "recursivo". Existen otras dos alternativas ("rolling" y "fixed") que se consideran en McCracken (2000), por ejemplo, pero nosotros no las contemplaremos.

[7] Los artículos de West (1996) y McCracken (2000) son más generales a este respecto, en el sentido de que f no tiene por qué representar una función de pérdidas. Pero éste es el significado que nos interesa en nuestro contexto.

esta sección, para tener una visión completa de la literatura en cuestión. Respecto a la teoría sobre modelos anidados, presentamos ahora una introducción breve, pero será en la sección 4 cuando se profundizará sobre ella.

2.1.1. Pérdidas diferenciables

West (1996) deriva la expresión correcta de la matriz de varianzas y covarianzas Ω para el estadístico $\overline{f} - E(f_t)$, bajo los siguientes supuestos:

a) Supuesto 1: $f_t(\beta)$ es dos veces diferenciable respecto a β en un entorno abierto de β^*.

b) Supuesto 2: La estimación $\widehat{\beta}_t$ verifica $\widehat{\beta}_t - \beta^* = B(t)H(t)$, donde $B(t)$ es $k \times q$ y $H(t)$ es $q \times 1$; $B(t) \xrightarrow{cs} B$, siendo B matriz de rango k; $H(t) = t^{-1}\sum_{s=1}^{t} h_s(\beta^*)$, para una $q \times 1$ condición de ortogonalidad $h_s(\beta^*)$ que cumple $E\left[h_s(\beta^*)\right] = 0$. Véase que, normalmente, este supuesto implica que el estimador utilizado sea consistente.

Por ejemplo, sea el modelo escalar $y_t = Z_t'\beta^* + u_t$ y sea la técnica de estimación OLS, se tiene que $B(t) = \left(\frac{1}{t}\sum_{s=1}^{t} Z_s Z_s'\right)^{-1}$, $B = E(Z_t Z_t')^{-1}$ y $h_s(\beta^*) = Z_s u_s$ (en este caso, $q = k$).

c) Supuesto 3: El vector $[vec(f_{t\beta})', f_t', h_t']'$ es estacionario en covarianza, siendo $f_{t\beta} = \left[\frac{\partial f_t}{\partial \beta}\right]_{\beta=\beta^*}$.[8]

d) Supuesto 4: $\lim_{T\to\infty} R, P = \infty$ y $\lim_{T\to\infty} P/R = \pi$, $0 \leq \pi \leq \infty$ ($\pi = \infty \Leftrightarrow \lim_{T\to\infty} R/P = 0$). Es decir, tanto R como P se hacen arbitrariamente grandes al crecer T, pero su ratio P/R converge a un número real no negativo.

Bajo los supuestos (1)-(4) y si S_{ff} es definida positiva,[9] West (1996) demuestra que

$$\sqrt{P}(\overline{f} - E(f_t)) \overset{a}{\sim} N(0, \Omega), \tag{1a}$$

$$\Omega = S_{ff} + \Pi(FBS_{fh}' + S_{fh}B'F') + 2\Pi F V_\beta F', \tag{1b}$$

siendo:

$$S_{ff} = \sum_{j=-\infty}^{+\infty} \Gamma_{ff}(j),\ \Gamma_{ff}(j) = E\left[(f_t - E(f_t))(f_{t-j} - E(f_t))'\right], \tag{2}$$

$$S_{fh} = \sum_{j=-\infty}^{+\infty} \Gamma_{fh}(j),\ \Gamma_{fh}(j) = E\left[(f_t - E(f_t))h_{t-j}'\right] \tag{3}$$

$$S_{hh} = \sum_{j=-\infty}^{+\infty} \Gamma_{hh}(j),\ \Gamma_{hh}(j) = E\left[h_t h_{t-j}'\right],\ V_\beta = BS_{hh}B',\ F = E\left[\frac{\partial f(\beta)}{\partial \beta}\right]_{\beta=\beta^*} \tag{4}$$

$$\Pi = 1 - \frac{\ln(1+\pi)}{\pi} \text{ si } 0 < \pi < \infty,\ \Pi = 0 \text{ si } \pi = 0 \text{ y } \Pi = 1 \text{ si } \pi = \infty. \tag{5}$$

Las matrices S_{ff}, S_{fh}, S_{hh}, F, B, V_β son de dimensiones $l \times l$, $l \times q$, $q \times q$, $l \times k$, $k \times q$, $k \times k$, respectivamente, y, con ello, Ω es $l \times l$. $V_\beta = BS_{hh}B'$ es la matriz asintótica de varianzas y covarianzas de $\sqrt{T}(\widehat{\beta}_T - \beta^*)$. Si no se está realizando estimación para generar las predicciones, entonces $B = 0$ y, por tanto, $\Omega = S_{ff}$. Ésa es una situación donde no existe IP. Otra situación diferente es aquella en la que *existe estimación paramétrica, pero es irrelevante* en términos de la matriz de varianzas y covarianzas del estadístico de contraste considerado por West (1996), situación que denotamos por IAIP. Esto ocurre si $\pi = 0$ ó si $F = 0$, escenarios en los que es obvio que $\Omega = S_{ff}$.[10] El primero de estos casos se refiere a situaciones en donde, para T arbitrariamente grande, el número de observaciones disponibles para la

[8] Los supuestos 1 y 3 de West (1996) incorporan además otras condiciones técnicas más complejas. Consúltese dicha referencia para detalles, págs 1070-1071.

[9] West (1996) exige además que una matriz S, cuyos bloques se corresponden con algunas de las matrices especificadas aquí, sea también definida positiva. Véase West (1996) para detalles, pág 1072.

[10] En realidad, existe una tercera condición de IAIP: $\Pi(FBS_{fh}' + S_{fh}'B'F') = -2\Pi F V_\beta F'$. Esta situación es muy poco frecuente. En West (1996), págs. 1073-4, se caracterizan algunos contextos predictivos que generan el cumplimiento de dicha condición.

estimación es arbitrariamente grande en relación al de previsiones, de modo que β^* puede tratarse como conocido. El segundo caso es de gran importancia en el desarrollo del capítulo y se puede alcanzar según el tipo de modelo de previsión y de función de pérdida. Tal y como muestra West (1996), una situación de este tipo se produce usando la función SE para evaluar previsiones de modelos no anidados y donde los errores de previsión y los regresores sean incorrelados. Por ejemplo, supóngase una variable y_t que se trata de prever a través de un modelo lineal $y_t = \varphi(Z_t, \beta^*)+u_t$, donde u_t es el error de previsión (que incluye la perturbación del verdadero PGD y el error de especificación). Sea $f_{t+1}(\beta^*) = (y_{t+1} - \varphi(Z_{t+1}, \beta^*))^2$ la función de pérdidas SE para horizonte $h = 1$, entonces $\left[\frac{\partial f(\beta)}{\partial \beta}\right]_{\beta=\beta^*} = -2u_{t+1}\nabla_{\beta^*}\varphi(Z_{t+1})$, donde $\nabla_{\beta^*}\varphi(Z_{t+1}) = \left[\frac{\partial \varphi(Z_t,\beta)}{\partial \beta}\right]_{\beta=\beta^*}$ y $F = -2E(u_{t+1}\nabla_{\beta^*}\varphi(Z_{t+1}))$, de modo que $F = 0 \Leftrightarrow E(u_{t+1}\nabla_{\beta^*}\varphi(Z_{t+1})) = 0$. Como se dijo en la introducción, una condición que garantiza la igualdad anterior es $E(u_{t+1}|Z_{t+1}) = 0$. Si el modelo fuera lineal ($\varphi(Z_t, \beta^*) = Z_t'\beta^*$), se tendría $F = 0 \Leftrightarrow E(u_{t+1}Z_{t+1}) = 0$, por lo que una condición análoga en dicho caso es que los errores y las variables explicativas estén incorrelados contemporáneamente.

La forma en la que West (1996) logra derivar la distribución asintótica de $\sqrt{P}(\overline{f} - E(f_t))$ es en base a una expansión de Taylor sobre $f_{t+h}(\widehat{\beta}_t)$ en el entorno de β^*, que permite distinguir entre la varianza del estadístico que es atribuible a la estimación paramétrica y el resto:

$$P^{1/2}(\overline{f} - E(f_t)) = P^{-1/2}\sum_{t=R}^{T}[f_{t+h}(\beta^*) - E(f_t)] + P^{-1/2}\sum_{t=R}^{T}\left[\frac{\partial f(\beta)}{\partial \beta}\right]_{\beta=\beta^*}(\widehat{\beta}_t - \beta^*) + o_p(1), \qquad (6)$$

donde $o_p(1)$ se refiere a un término que, bajo condiciones muy generales, es irrelevante asintóticamente. A partir de la expansión (6) y usando los supuestos mencionados antes, West (1996) obtiene la siguiente expresión:

$$P^{1/2}(\overline{f} - E(f_t)) = P^{-1/2}\sum_{t=R}^{T}[f_{t+h}(\beta^*) - E(f_t)] + P^{-1/2}FB\sum_{t=R}^{T}H(t) + o_p(1), \qquad (7)$$

Finalmente, West (1996) deriva la distribución conjunta de $P^{-1/2}\left(\sum[f_{t+h}(\beta^*) - E(f_t)], B\sum H(t)\right)$, y aplicándola sobre (7), se tiene la del estadístico de contraste $P^{1/2}(\overline{f} - E(f_t))$. Hemos mostrado brevemente este proceso por dos motivos:

1) Fundamentalmente, porque tanto (6) como (7) permiten distinguir claramente las fuentes de varianza en el estadístico $P^{1/2}(\overline{f} - E(f_t))$. El primer sumando es el que genera una varianza que procede del hecho de que f_{t+h} sea una variable aleatoria, es decir, dicha varianza existiría incluso si β^* fuera conocido. Esta varianza es S_{ff} y es la única que se tiene en cuenta en la versión original del test de Diebold y Mariano (1995), por ejemplo, y la única que hemos considerado nosotros en los capítulos anteriores. Por su parte, el segundo sumando genera la varianza causada por la estimación paramétrica, los dos últimos términos de Ω en (1b).

2) En segundo lugar, porque nos servirá en la sección 4 para entender la razón por la que ni el resultado de West (1996) ni el de Diebold y Mariano (1995) aplican cuando el contraste compara dos modelos anidados.

2.1.2. Pérdidas no diferenciables

McCracken (2000) extiende el teorema de West (1996) para funciones de pérdida no diferenciables. Las modificaciones que expone son, esencialmente, las siguientes:

a) Supuesto 1: Sea $g_t = [f_t(\beta) - E(f_t(\beta))', h_t'(\beta)]'$, se exige diferenciabilidad respecto a β en un entorno de β^* para $E(g_t(\beta))$, en vez de para $f_t(\beta)$.

b) Supuesto 3: Se exige estacionariedad en covarianza respecto al vector aleatorio $g_t(\beta^*)$, en vez de respecto a $[vec(f_{t\beta})', f_t', h_t']'$.

Los Supuestos 2 y 4 se mantienen inalterados en relación al trabajo de West (1996).

c) El enunciado del teorema continúa siendo el especificado en (1a), pero ahora el significado de la matriz F pasa a ser $F = \left[\frac{\partial E f(\beta)}{\partial \beta}\right]_{\beta=\beta^*}$, en vez de $F = E\left[\frac{\partial f(\beta)}{\partial \beta}\right]_{\beta=\beta^*}$.

Éste es el teorema que aplica para nuestra función de pérdida discreta. En lo que sigue, y para contextos de previsión que verifiquen los Supuestos 2, 3 y 4 anteriores, trataremos primero de demostrar que $E(f_t(\beta))$

es diferenciable en β, siendo f nuestra función discreta, y, una vez hecho esto, se intentará obtener una expresión analítica para $F = \left[\frac{\partial Ef(\beta)}{\partial \beta}\right]_{\beta=\beta^*}$.

McCracken (2000) (págs. 204-5) muestra la expresión que toma F en modelos lineales no anidados bajo una función de pérdidas no diferenciable típica, el AE: $F = -E\left[sign(u_{t+1})Z'_{t+1}\right]$, siendo $sign(x)$ una función que toma valor -1 si $x < 0$ y 1 si $x \geq 0$. Por lo tanto, $F = 0$ en este caso si los errores u_{t+1} y los regresores Z_{t+1} están incorrelados y, además, la distribución de probabilidad de u_{t+1} es simétrica. Véase que si se diera la circunstancia de que $F \neq 0$ usando el AE como pérdida, su estimación sería sencilla, por la simplicidad de la expresión analítica correspondiente, pero, desgraciadamente, este comentario no es, ni mucho menos, extensible a la mayor parte de funciones no diferenciables.

2.1.3. Modelos anidados

Cuando los modelos son anidados, el teorema de West (1996) no es válido, ni tampoco el contraste Diebold y Mariano (1995) lo es de forma general, incluso en un marco de no estimación paramétrica. En el apartado 4.1 se presenta una argumentación para justificar esta afirmación. En dicho contexto, el resultado de mayor interés se encuentra en McCracken (2007), donde se deriva la distribución asintótica para un contraste concreto de comparación de habilidad predictiva entre dos modelos (OOS-t test), que es el mismo test DM pero particularizado al caso en el que el diferencial de pérdidas $d_t = f_{1t} - f_{2t}$ no presente autocorrelación. En dicha situación, y bajo los supuestos de diferenciabilidad de $f_t(\beta)$ y de que la función asociada al método de estimación sea la misma que define las pérdidas (normalmente, OLS y SE), McCracken (2007) obtiene la distribución asintótica, que ya no estándar, correspondiente al estadístico $\sqrt{P}(\overline{f}_1 - \overline{f}_2)\widehat{\Omega}_d^{-1/2}$, donde $\widehat{\Omega}_d$ debe ser $\widehat{\Omega}_d = P^{-1}\sum_{t=R}^{T}(f_{1t}(\widehat{\beta}_{1t}) - f_{2t}(\widehat{\beta}_{2t}))^2$. En este caso, el estadístico de contraste no necesita incorporar correcciones por IP, pero, a cambio, la distribución resultante depende del grado de IP, medida ésta a través del parámetro π. Véase la referencia citada para detalles. No existe literatura análoga sobre modelos anidados adaptada al caso de funciones de pérdida no diferenciables.

2.2. Nota para aplicar los resultados sobre IP en los tests de comparación de capacidad predictiva

Nuestro objetivo es comprobar cómo afecta la IP al tamaño de tests que emplean la distribución asintótica del estadístico $\overline{f} - E(f_t)$ cuando f_t representa una función de pérdidas discreta. Los contrastes que nos interesan a este respecto son el test C1, en sus dos versiones (Capítulo 1), respecto a evaluación de previsiones de un modelo, y el test DM implementado con función discreta (Capítulo 2), respecto a comparación entre dos alternativos. Los resultados que se obtengan para DM son igualmente válidos para la versión aproximada del test Mult2 (Mult2-aprx), presentada en aquel capítulo.

Nos centraremos siempre en el horizonte de previsión uno ($h = 1$).

Para tests de comparación de capacidad predictiva entre dos modelos, se tendrá, por tanto, que $l = 2$, $f_t = (f_{1t}, f_{2t})'$ y que las funciones de pérdidas f_{1t} y f_{2t} serán las mismas para ambos modelos. Como los contrastes que manejamos de este tipo emplean el estadístico $\overline{f}_1 - \overline{f}_2 - E(f_1 - f_2)$, el enunciado de los teoremas de West (1996) y McCracken (2000) se puede modificar fácilmente al caso que nos ocupa del siguiente modo:

$$\begin{gathered}\sqrt{P}(\overline{f}_1 - \overline{f}_2 - E(f_1 - f_2)) \overset{a}{\sim} N(0, \Omega^{(2)}), \\ \Omega^{(2)} = \Omega_{11} + \Omega_{22} - 2\Omega_{12},\end{gathered} \tag{8}$$

siendo Ω_{ij} el elemento i, j de la matriz Ω 2×2 definida en los teoremas anteriores. Por tanto, $\Omega^{(2)}$ es la varianza asintótica correcta bajo IP para el test DM, mientras $S_{ff}^{(2)} = S_{ff\ 11} + S_{ff\ 22} - 2S_{ff\ 12}$ es la varianza empleada habitualmente (por ejemplo, en el Capítulo 2), que se derivaba suponiendo β^* conocido.

2.3. Repaso de la definición de la función de pérdidas discreta

Aunque ya ha sido explicado en capítulos anteriores, recordaremos la definición que utilizamos para la función de pérdidas a la que nos estamos refiriendo con el adjetivo "discreta". Sea y_t el dato en el periodo t y $v_{i,t}$ la previsión (realizada en $t-h$) para y_t por el modelo de previsión i-ésimo ($i = 1, 2$), la función discreta f asigna al par $(y_t, v_{i,t})$ una pérdida de entre un conjunto de valores A finito (y, normalmente, pequeño).[11] Formalmente, el modo de definir f es el que se describe a continuación:

a) Se establece una partición P del dominio de datos (y previsiones), es decir, de la recta real R, en n regiones: $P = \{l_0, l_1, l_2, ..., l_{n-2}, l_{n-1}, l_n\}$, siendo $l_0 = -\infty$ y $l_n = +\infty$. Denotaremos las regiones por r_1, r_2, ..., r_n. Implícitamente, se ha particionado el dominio $R^2 = R \times R$ en n^2 "cuadrantes".

b) Se asigna un valor del conjunto A a cada cuadrante (j, k). Es decir, $f(y_t, v_{i,t}) = a_{jk} \Leftrightarrow l_{j-1} \leq y_t \leq l_j$ y $l_{k-1} \leq v_{i,t} \leq l_k$. Por lo tanto, la función de pérdidas discreta f puede especificarse por una matriz cuya entrada (j, k) (fila j y columna k) representa la pérdida correspondiente al cuadrante $y_t \in r_j = (l_{j-1}, l_j)$, $v_{i,t} \in r_k = (l_{k-1}, l_k)$.

Por el momento, solo supondremos que los valores en A son numéricos y que el valor asignado a los cuadrantes (i, i) es cero.

Un ejemplo sencillo y razonable de la función f puede ser el siguiente:

		$v_{i,t}$			
		G−	P−	P+	G+
y_t	G−	0	1	2	3
	P−	1	0	2	3
	P+	3	2	0	1
	G+	3	2	1	0

(9)

, donde G−, P−, G+, P+ simbolizan valores "negativos/positivos y grandes/pequeños", respondiendo a una partición $P = \{-\infty, -q, 0, +q, +\infty\}$, por ejemplo, con q cualquier valor real adecuado al caso.

El interés conceptual de este tipo de funciones puede consultarse en la Introducción de la Tesis, y en las secciones de Introducción de los Capítulos 1 y 2.

[11] Recordemos de nuevo que, aunque en el capítulo anterior esta función se denotaba por g, ahora pasaremos a usar la letra f para designarla, por analogía con la notación de West (1996) y McCracken (2000).

3. Modelos lineales no anidados (variables estacionarias)

Pretendemos analizar el impacto de la IP en tests que usan la formulación (1a), en general, o una transformación de ella (por ejemplo, (8) en tests como DM, comparando capacidad predictiva de $l = 2$ modelos) bajo una función de pérdidas f de tipo discreto, cuando los l modelos de previsión son lineales y todas las variables involucradas en ellos, estacionarias en sentido débil. Seguiremos el esquema de trabajo anticipado en la Introducción:

Como paso primero y fundamental, trataremos de hallar una expresión para F, respondiendo a la definición de McCracken (2000), que es la que aplica a nuestra f. Este paso se corresponde con el objetivo a) citado en la Introducción. Sin embargo, el teorema de McCracken (2000) − igual que el de West (1996) − solo es válido si no hay dos modelos anidados. Por lo tanto, la situación que se excluye en esta sección es aquella en la que los regresores de un modelo de previsión sean exactamente los mismos que los de otro, más un grupo de regresores cuyos parámetros asociados son nulos. Dicho formalmente, en esta sección se excluye, por tanto, el caso $\beta_i^* = (\beta_j^*, 0_{k_i - k_j})$, donde β_s^* representa el verdadero valor del vector paramétrico correspondiente al modelo M_s, y siendo $k_i - k_j > 0$.[12]

Una vez se tenga una expresión para F, el segundo paso será caracterizar qué casos implican $F = 0$. Éste es el propósito b) especificado en la Introducción.

Como veremos más adelante, bajo no independencia entre errores de previsión y regresores, ocurrirá que $F \neq 0$, de modo que ignorar la IP genera sesgo asintótico en el tamaño de los tests. Pues bien, el último paso será desarrollar el punto c) de la Introducción: calcular numéricamente dicho sesgo para un ejercicio de simulación representativo, bajo tres funciones de pérdida alternativas: SE, AE y una función discreta. Nos centraremos en el caso $l = 2$ y en el test DM (cuya formulación será (8)). El ejemplo elegido será exactamente el presentado por West (1996) (págs 1076-9).

3.1. Contexto general

Sea M_i un modelo de previsión $y_t = Z_{it}'\beta_i^* + u_{it}$, $i = 1, ..., l$, que prevé y_{t+1} por $v_{i,t+1} = Z_{it+1}'\beta_i$, denotando por β_i la estimación del vector β_i^* $(K_i \times 1)$ con las observaciones muestrales hasta t, y por u_{it+1} el error de previsión teórico para M_i.[13] Como ya se dijo arriba, no se admiten dos modelos anidados. Los vectores β_i se agrupan en β y sus verdaderos valores en β^*, de tamaño $K \times 1$. Por su parte, el vector de regresores Z_{it} se compone, para todo i, de variables estacionarias en sentido débil, con varianza finita.

Siguiendo lo expuesto en el apartado 2.3, la función de pérdidas discreta se define por:

$$f_{it+1}(\beta) = a_{jk},\ l_{j-1} \leq y_{t+1} \leq l_j,\ l_{k-1} \leq v_{i,t+1} \leq l_k. \tag{10}$$

Aplicando McCracken (2000), la definición de F que procede es $F = \left[\frac{\partial E f(\beta)}{\partial \beta}\right]_{\beta=\beta^*}$. Calcularemos el elemento genérico $F_{i,w} = \left[\frac{\partial E f_{it+1}(\beta)}{\partial \beta(w)}\right]_{\beta=\beta^*}$, para cualquier modelo M_i y cualquier parámetro β_w. Si $\beta(w) \notin \beta_i$, es decir, si el elemento w-ésimo de β no aparece en la especificación de M_i, la derivada será cero, obviamente. Nos concentraremos en las derivadas respecto al resto de parámetros. Con intención de evitar que la exposición sea innecesariamente farragosa, nos permitiremos la licencia de denotar por β_1 el parámetro respecto al que se deriva, y por β_2 el vector de dimensión $K_i - 1$, con el resto de parámetros del modelo M_i, y lo mismo haremos para los regresores asociados. Fijémonos que eliminamos también el subíndice que define el modelo. Así, el modelo original $y_t = Z_{it}'(w)\beta^*(w) + Z_{it}^{e\prime}\beta_i^{e*} + u_{it}$ (donde β_i^{e*} se refiere a todos los parámetros del modelos salvo $\beta^*(w)$, y donde Z_{it}^e se define análogamente) pasa a escribirse como $y_t = z_{1t}\beta_1^* + Z_{2t}'\beta_2^* + u_t$.

[12] La definición de modelos anidados es más general que la aquí sugerida (véase definición formal de Davidson y MacKinnon (2004), "Econometric Theory and Methods", página 669), pero, en el contexto que nos ocupa ahora, basta con pensar en este caso.

[13] Se utiliza la notación β en vez de $\widehat{\beta}$ por coherencia con los artículos de West (1996) y McCracken (2000).

Tras esto, la definición (10) de f queda del siguiente modo:

$$\begin{aligned} f_{it+1}(\beta) = a_{jk},\ l_{j-1} \leq z_{1t+1}\beta_1^* + Z'_{2t+1}\beta_2^* + u_{t+1} \leq l_j,\ l_{k-1} \leq z_{1t+1}\beta_1 + Z'_{2t+1}\beta_2 \leq l_k \\ \Leftrightarrow b_{j-1} = l_{j-1} - z_{1t+1}\beta_1^* - Z'_{2t+1}\beta_2^* \leq u_{t+1} \leq l_j - z_{1t+1}\beta_1^* - Z'_{2t+1}\beta_2^* = b_j \\ \text{y } d_{k-1} = \frac{l_{k-1} - Z'_{2t+1}\beta_2}{\beta_1} \leq z_{1t+1} \leq \frac{l_k - Z'_{2t+1}\beta_2}{\beta_1} = d_k. \end{aligned} \tag{11}$$

Los parámetros b_{j-1}, b_j, d_{k-1} y d_k se han definido para facilitar el desarrollo a continuación.

3.2. Expresión general de la matriz F, bajo función de pérdidas discreta

3.2.1. Expresión analítica para $E(f_{t+1}(\beta))$

Se utilizará la notación $g_x(x)$ para designar la función de densidad marginal (si x escalar) o conjunta (si x vectorial) de x, $g_{x_1|x_2}(x_1|x_2)$ para la función de densidad de x_1 condicionada a x_2, $G_x(x)$ para la función de distribución de x y $G_{x_1|x_2}(x_1|x_2)$ para la función de distribución de x_1 condicionada a x_2. Además, en lo que sigue, eliminaremos los subíndices temporales de las variables para facilitar la exposición:

$$E(f_{t+1}(\beta)) = \sum_{k=1}^{n}\sum_{j=1}^{n} a_{jk} P(l_{j-1} \leq y_{t+1} \leq l_j, l_{k-1} \leq v_{t+1} \leq l_k) =$$

$$= \sum_{k=1}^{n}\sum_{j=1}^{n} a_{jk} \int_{Z_2} \left[\int_{d_{k-1}}^{d_k} \left[\int_{b_{j-1}}^{b_j} g_{u|z_1,Z_2}(u|z_1,Z_2)du \right] g_{z_1|Z_2}(z_1|Z_2)dz_1 \right] g_{Z_2}(Z_2)dZ_2.$$

Las variables b_{j-1}, b_j, d_{k-1}, d_k se definieron en (11). Téngase presente que b_{j-1} y b_j dependen de z_1 y Z_2, y d_{k-1} y d_k dependen de Z_2, β_1 y β_2. Aplicando la definición de función de distribución condicional, se tiene:

$$E(f_{t+1}(\beta)) = \sum_{k=1}^{n}\sum_{j=1}^{n} a_{jk} \int_{Z_2} \left[\int_{d_{k-1}}^{d_k} \left(G_{u|z_1,Z_2}(b_j|z_1,Z_2) - G_{u|z_1,Z_2}(b_{j-1}|z_1,Z_2) \right) g_{z_1|Z_2}(z_1|Z_2)dz_1 \right] g_{Z_2}(Z_2)dZ_2.$$

Véase que $E(f_{t+1}(\beta))$ es continua en β siempre que el error de previsión u sea una variable aleatoria continua. Bajo dicho supuesto, en principio poco restrictivo, el teorema de McCracken (2000) es aplicable en nuestro contexto.

3.2.2. Expresión analítica para un elemento genérico de F

Ya estamos en disposición de calcular el elemento genérico de la matriz F, es decir, de calcular $\left[\frac{\partial E f(\beta)}{\partial \beta_1}\right]_{\beta=\beta^*}$. Podemos escribir $E(f_{t+1}(\beta))$ como $E(f_{t+1}(\beta)) = \sum_{k=1}^{n}\sum_{j=1}^{n} a_{jk} S(j,k)$, y calcular la derivada para el sumando $S(j,k)$. Para ello, aplicaremos la regla de Leibniz:

$$\frac{\partial}{\partial x} \int_{a(x,s)}^{b(x,s)} h(y,s)dy = h(b(x,s),s)\frac{\partial b}{\partial x} - h(a(x,s),s)\frac{\partial a}{\partial x}.$$

Véase que, en nuestro caso, $a = d_{k-1}$, $b = d_k$, $x = \beta_1$, $y = z_1$, $s = Z_2$. La derivada de $S(j,k)$ es:

$$\frac{\partial}{\partial\beta_1}S(j,k)=\int_{Z_2}\left(G_{u|z_1,Z_2}(l_j-d_k\beta_1^*-Z_2'\beta_2^*|d_k,Z_2)-G_{u|z_1,Z_2}(l_{j-1}-d_k\beta_1^*-Z_2'\beta_2^*|d_k,Z_2)\right)$$
$$g_{z_1|Z_2}(d_k|Z_2)\frac{\partial d_k}{\partial\beta_1}g_{Z_2}(Z_2)dZ_2-$$
$$-\int_{Z_2}\left(G_{u|z_1,Z_2}(l_j-d_{k-1}\beta_1^*-Z_2'\beta_2^*|d_{k-1},Z_2)-G_{u|z_1,Z_2}(l_{j-1}-d_{k-1}\beta_1^*-Z_2'\beta_2^*|d_{k-1},Z_2)\right)$$
$$g_{z_1|Z_2}(d_{k-1}|Z_2)\frac{\partial d_{k-1}}{\partial\beta_1}g_{Z_2}(Z_2)dZ_2.$$

Dado que, tras evaluar $d_k^*=d_k(\beta^*)$, resulta $l_j-d_k(\beta^*)\beta_1^*-Z_2'\beta_2^*=l_j-\frac{l_k-Z_2'\beta_2^*}{\beta_1^*}\beta_1^*-Z_2'\beta_2^*=l_j-l_k$ y que $\left[\frac{\partial d_k}{\partial\beta_1}\right]_{\beta=\beta^*}=-\frac{l_k-Z_2'\beta_2^*}{(\beta_1^*)^2}=-\frac{d_k^*}{\beta_1^*}$, se obtiene inmediatamente el siguiente resultado:

$$\left[\frac{\partial}{\partial\beta_1}S(j,k)\right]_{\beta=\beta^*}=-\int_{Z_2}\left[B(j,k)H_k-B(j,k-1)H_{k-1}\right]g_{Z_2}(Z_2)dZ_2,$$
$$H_k=g_{z_1|Z_2}(d_k^*|Z_2)\frac{d_k^*}{\beta_1^*},\ B(j,k)=G(j,k)-G(j-1,k)\text{ y }G(j,k)=G_{u|z_1,Z_2}(l_j-l_k|d_k^*,Z_2).\qquad(12)$$

Por tanto, el elemento genérico $F_{i,w}$ de la matriz F será:

$$F_{i,w}=-E[R(Z_2)]=-\int_{Z_2}R(Z_2)g_{Z_2}(Z_2)dZ_2,$$
$$\text{siendo }R(Z_2)=\sum_{k=1}^{n}\sum_{j=1}^{n}a_{jk}\varphi_{jk}\text{ y }\varphi_{jk}(Z_2)=\left[B(j,k)H_k-B(j,k-1)H_{k-1}\right].$$

Fijémonos que φ_{jk} es una función del vector aleatorio Z_2 y de los parámetros l_j, l_k y β^*.

Reordenando términos:

$$R(Z_2)=\left[\sum_{k=1}^{n-1}H_k\left(\sum_{j=1}^{n}B(j,k)(a_{jk}-a_{jk+1})\right)\right]+H_n\sum_{j=1}^{n}B(j,n)a_{jn}-H_0\sum_{j=1}^{n}B(j,0)a_{j1}.$$

Finalmente, $l_0=-\infty$ y $l_n=+\infty$ (véase apartado 2.3)$\Rightarrow d_0^*=-\infty$ y $d_n^*=+\infty$ y, en consecuencia, se tiene también que $g_{z_1|Z_2}(d_0^*|Z_2)=g_{z_1|Z_2}(d_n^*|Z_2)=0$. Dado que se supuso varianza finita para cualquier regresor, $H_n=\lim_{k\to+\infty}H_k=g_{z_1|Z_2}(k|Z_2)\frac{k}{\beta_1^*}=0$ y $H_0=\lim_{k\to-\infty}H_k=g_{z_1|Z_2}(k|Z_2)\frac{k}{\beta_1^*}=0$.[14]

Por lo tanto:

$$F_{i,w}=-E[R(Z_2)]=-\int_{Z_2}R(Z_2)g_{Z_2}(Z_2)dZ_2=-\int_{Z_2}\left(\sum_{k=1}^{n-1}H_kQ_k\right)g_{Z_2}(Z_2)dZ_2,\qquad(13)$$
$$\text{siendo }Q_k=\sum_{j=1}^{n}B(j,k)(a_{jk}-a_{jk+1}),$$

y donde $B(j,k)$, H_k y $d_k(\beta)$ se pueden encontrar en (12), (12) y (11), respectivamente

Recuérdense las equivalencias del principio de apartado: $z_{1t}=Z_{it}'(w)$, $\beta_1=\beta(w)$, $Z_{2t}'=Z_{it}^{e\prime}$, $\beta_2=\beta_i^e$ y $u_t=u_{it}$.

La formulación (13) constituye la expresión analítica general de la matriz F bajo f discreta. No obstante, todavía no se ha supuesto nada en particular sobre la especificación de la función de pérdida discreta. A continuación, se enumerarán y justificarán una serie de supuestos razonables sobre ella, bajo los cuales se obtienen expresiones más simplificadas de F y que permitirán obtener en el apartado 3.4 conclusiones relevantes sobre el valor de dicha matriz en la práctica.

[14] La finitud de la varianza de la distribución de $z_1|Z_2$ garantiza que la convergencia hacia cero de la densidad de dicha variable es más rápida que lineal, y, por tanto, el límite del producto $g_{z_1|Z_2}(k|Z_2)k=0$ es cero.

3.3. Supuestos sobre la estructura de la función de pérdidas discreta

A continuación caracterizamos cierto tipo de función de pérdidas discreta, a través de cuatro enunciados. Se supondrá que tres de ellos (1, 3 y 4) se están cumpliendo en todo momento a lo largo del desarrollo de las expresiones en el próximo apartado. El otro solo es necesario en un punto final de la demostración.

Separamos los supuestos según se refieran al modo en que debe llevarse a cabo la partición del dominio de datos y previsiones o a la forma de asignar los valores numéricos de las pérdidas en dicha partición (fases a) y b) de la definición del apartado 2.3).

Sobre la especificación de la partición :

1) Supuesto 1: $l_s = 0$, $1 \leq s \leq n-1$.

Es decir, la clasificación de datos y previsiones utiliza una partición base que separa valores positivos y negativos, y, dentro de cada una de estas zonas, se separará en regiones. Éste es un supuesto que no es determinante en muchos de los resultados que se expondrán, pero que es necesario para que los enunciados de los siguientes supuestos cobren sentido.

Aunque pudiera parecerlo, no es un supuesto demasiado restrictivo. Conceptualmente, nuestra función de pérdidas está construida para casos de este tipo, donde interesa clasificar las previsiones primero según signo y, después, según cuantía (para ejemplos y argumentos a favor de este tipo de clasificación, véase el Capítulo 1 de esta tesis). Esta forma de clasificar se puede generalizar, en realidad, a situaciones donde interesa particionar en torno a un valor $\overline{y}$, en cuyo caso los datos y previsiones a considerar serían $y_{t+1} - \overline{y}$ y $v_{i,t+1} - \overline{y}$, respectivamente, y l_s seguiría siendo 0 en vez de $\overline{y}$.

2) Supuesto 2: $s = n/2$, $l_{s-k} = -l_{s+k}$, $k = 1, ..., s$.

Es decir, la partición es simétrica respecto a $l_s = 0$. Como ya se dijo arriba, este supuesto no juega papel en la obtención de varios resultados intermedios del apartado 3.4, pero sí para el resultado fundamental (Corolario 1 del apartado 3.4). En realidad, para lograr dicho resultado, bastaría una versión restringida del Supuesto 2, que denotaremos por Supuesto 2a:

Supuesto 2a: $l_{s-1} = -l_{s+1}$. Sin embargo, conceptualmente, parece más razonable la versión general anterior, por lo que será la que utilizaremos en adelante.

Sobre la asignación de pérdidas a_{jk} a la partición:

3) Supuesto 3: Sea $k \neq s$, entonces $a_{jk+1} - a_{jk} = \begin{array}{l} +c,\ k \geq j \\ -c,\ k < j \end{array}$.

Es decir, una vez el dato y cae en una región de la partición r_j, las penalizaciones asociadas a previsiones del mismo signo se incrementan siempre una cuantía fija c conforme la región r_k a la que corresponden se aleja de la región correcta. Por tanto, previsiones del mismo signo tienen asignada una pérdida que crece linealmente con el error en cuantía, medido éste por $|k - j|$. Este supuesto es razonable en muchas aplicaciones, pero no en todas.[15]. Las matrices (9), en el apartado 2.3, y (14), a continuación en el apartado actual, representan funciones de pérdida que verifican este supuesto (en ambos casos, se cumple que $c = 1$).

4) Supuesto 4: 4.a) $a_{js+1} - a_{js} = \begin{array}{l} +c_s,\ s > j \\ -c_s,\ s+1 < j \end{array}$; 4.b) $a_{js+1} - a_{js} = \begin{array}{l} +c'_s,\ s = j \\ -c'_s,\ s+1 = j \end{array}$.

El Supuesto 4 es del mismo tipo a 3, solo que se refiere a las columnas centrales de la matriz de pérdidas. La razón de distinguir entre columnas centrales (sobre las que aplica 4) y no centrales (sobre las que aplica 3) es que en las primeras es donde se establecen los aumentos de pérdidas entre previsiones de signo correcto e incorrecto. Obviamente, es razonable permitir que dichos cambios sean de una cuantía en general diferente a (en principio, mayor que) en otras zonas de la matriz. Ésta es la razón de ser del Supuesto 4.

Sin embargo, es muy posible que el usuario necesitara, para generar una matriz de pérdidas razonable, utilizar cambios entre las dos *columnas* centrales que sean diferentes según se encuentre en las *filas* centrales o en el resto, por razones que a continuación se entenderán. Éste es el motivo de desglosar 4 en 4.a y 4.b. Antes de dar paso a un ejemplo que permita intuir la razón de tal desglose, debemos introducir un principio razonable sobre la función de pérdidas (denotemos el principio por P), que *podría* ser deseado por el usuario de una función de pérdidas discreta: "todas las pérdidas asociadas a previsiones cuyo signo fue correcto deben ser menores que cualquiera de las pérdidas asociadas a previsiones que hayan errado el signo". Entiéndase que P es una norma que el usuario podría desear respetar a la hora de construir su matriz

[15] En algunos contextos de previsión, existen errores cuya cuantía es inaceptable. En esos casos, al pasar la previsión de cierto límite l_k, la pérdida puede aumentar más que linealmente.

de pérdidas, no un supuesto que nosotros establezcamos. El asunto es que, si nosotros queremos que los supuestos exigidos en la función de pérdidas no hagan imposible que se verifique P cuando lo desee el usuario, debemos permitir cambios en las pérdidas adyacentes de las columnas centrales (las que regulan el paso de previsiones de signo correcto a previsiones de signo incorrecto) de cuantía distinta en general según se esté en las filas centrales (cambios de cuantía c'_s) o no (cambios de cuantía c_s).

Para entenderlo, observemos lo que ocurre en la matriz (14) que hemos escrito abajo. Tomemos, por ejemplo, la fila 2. Las 3 primeras columnas se refieren a pérdidas para previsiones con signo correcto. La mayor pérdida asignada es 1 (cuadrantes (2,1) y (2,3)). Las columnas 4 a 6 contienen las pérdidas para previsiones de signo erróneo. Si quisiéramos respetar el principio P, debemos hacer que la pérdida del cuadrante (2,4) sea, al menos, igual a 2 (los cuadrantes (2,5) y (2,6) tendrán pérdidas aún mayores, lógicamente, ya que se corresponden con regiones donde el error es de magnitud mayor a la de la región asociada a (2,4)). Ésta es, precisamente, la pérdida que se está utilizando en nuestra matriz. Por lo tanto, se ha empleado $c_s = 1$. Bien, ahora tomemos la fila 3 (una de las dos centrales, por tanto). La mayor pérdida de las que corresponden a previsiones de signo correcto es 2 (en el cuadrante (3,1)). Por tanto, si queremos respetar P, necesitamos que la pérdida en (3,4) sea, al menos, 3. Por lo tanto, hemos utilizado $c'_s = 3 > c_s = 1$. Si nuestro Supuesto 4 no se desglosara en 4.a y 4.b, sino que las cuantías de los cambios de pérdidas en las columnas centrales de la matriz tuvieran que ser iguales en todas las filas (de cuantía c_s), hubiéramos tenido que asignar una pérdida igual a 1 en el cuadrante (3,4), violando el principio P.

Queremos dejar constancia de que el resultado fundamental que estableceremos próximamente en el Corolario 1 del apartado 3.4 se hubiera obtenido igualmente si hubiéramos eliminado el Supuesto 4 y extendido 3 también para $k = s$ (o si se prescinde de 4.b y se extiende 4.a para $j = s$ y $s + 1$), solo que, entonces, el resultado sería válido solo para matrices de pérdida con estructuras de menor riqueza (por ejemplo, P podría no estar garantizado) que aquellas para las que están diseñados nuestros Supuestos 3, 4.a.y 4.b.

		v					
		G−	M−	P−	P+	M+	G+
y	G−	0	1	2	3	4	5
	M−	1	0	1	2	3	4
	P−	2	1	0	3	4	5
	P+	5	4	3	0	1	2
	M+	4	3	2	1	0	1
	G+	5	4	3	2	1	0

(14)

Aunque los Supuestos 3 y 4 son razonables, es obvio que implican restringir de modo significativo el tipo de funciones de pérdida a manejar. Por ello, en el apartado 3.5.2 trataremos de ilustrar acerca de la magnitud de la distorsión que se produce en los resultados expuestos a continuación cuando se incumplen dichos supuestos.

3.4. Corolarios. Obtención analítica de la propiedad IAIP bajo función de pérdida discreta

En primer lugar, desarrollaremos la fórmula general (13) bajo el cumplimiento de los Supuestos 1, 3 y 4 presentados en el apartado anterior. Posteriormente, la expresión obtenida quedará simplificada al introducir hipótesis adicionales, esencialmente sobre la función de distribución $G_{u|z_1,Z_2}$, dando lugar a los resultados más relevantes de la sección.

[Lema 1]

Bajo el contexto definido en el apartado 3.1 (modelos de previsión lineales no anidados, con variables estacionarias) y una función de pérdidas discreta verificando los Supuestos 1, 3 y 4 del apartado 3.3, el componente $F_{i,w}$ de la matriz F de McCracken (2000) toma la expresión (13) con:

$$\begin{aligned} R(Z_2) &= H_s Q_s + c \sum_{k\in[1,n-1],\ k\neq s} H_k(1 - 2G_{u|z_1,Z_2}(0|d_k^*, Z_2)), \\ Q_s &= c_s + \left(G_{u|z_1,Z_2}(l_{s-1}|d_s^*, Z_2) + G_{u|z_1,Z_2}(l_{s+1}|d_s^*, Z_2)\right)(c'_s - c_s) - 2c'_s G_{u|z_1,Z_2}(0|d_s^*, Z_2) \end{aligned} \tag{15}$$

Demostración:

a) Sea $k \neq s$:

$Q_k = \sum_{j=1}^{n} B(j,k)(a_{jk} - a_{jk+1}) = \sum_{j=1}^{k} B(j,k)(a_{jk} - a_{jk+1}) + \sum_{j=k+1}^{n} B(j,k)(a_{jk} - a_{jk+1}) =$

$= c\left(\sum_{j=k+1}^{n} B(j,k) - \sum_{j=1}^{k} B(j,k) \right)$, donde la última igualdad se obtiene al aplicar el Supuesto 3.

Ahora, usando la definición de (12) $B(j,k) = G(j,k) - G(j-1,k)$, se tiene:

$Q_k = c[(G(k+1,k) - G(k,k) + G(k+2,k) - G(k+1,k) + ... + G(n,k) - G(n-1,k)) -$

$-(G(1,k) - G(0,k) + G(2,k) - G(1,k) + ... + G(k,k) - G(k-1,k))] =$

$= c\,(G(n,k) - G(k,k) + G(0,k) - G(k,k)) = c(1 - 2G_{u|z_1,Z_2}(0|d_k^*, Z_2))$, ya que $G_{u|z_1,Z_2}(-\infty|d_k^*, Z_2) = 0$ y $G_{u|z_1,Z_2}(+\infty|d_k^*, Z_2) = 1$, por definición de función de distribución de probabilidad.

b) Sea $k = s$:

$Q_s = \sum_{j=1}^{n} B(j,s)(a_{js} - a_{js+1}) = \sum_{j=1}^{s-1} B(j,s)(a_{js} - a_{js+1}) + \sum_{j=s+2}^{n} B(j,s)(a_{js} - a_{js+1}) +$

$+B(s,s)(a_{ss} - a_{ss+1}) + B(s+1,s)(a_{s+1s} - a_{s+1s+1}) =$

$= -\sum_{j=1}^{s-1} B(j,s)c_s + \sum_{j=s+2}^{n} B(j,s)c_s - B(s,s)c'_s + B(s+1,s)c'_s$, habiéndose aplicado el Supuesto 4 para obtener la última igualdad.

Procediendo de nuevo como en a), se tiene:

$Q_s = -c_s[(G(1,s) - G(0,s) + G(2,s) - G(1,s) + ... + G(s-1,s) - G(s-2,s)) -$

$(G(s+2,s) - G(s+1,s) + G(s+3,s) - G(s+2,s) + ... + G(n,s) - G(n-1,s))] +$

$+c'_s[-G(s,s) + G(s-1,s) + G(s+1,s) - G(s,s)] =$

$-c_s[G(s-1,s) - G(0,s) + G(s+1,s) - G(n,s)] + c'_s[G(s-1,s) - 2G(s,s) + G(s+1,s)] =$

$= (c'_s - c_s)[G(s-1,s) + G(s+1,s)] + c_s - 2c'_s G(s,s) =$

$c_s + \left(G_{u|z_1,Z_2}(l_{s-1}|d_s^*, Z_2) + G_{u|z_1,Z_2}(l_{s+1}|d_s^*, Z_2)\right)(c'_s - c_s) - 2c'_s G_{u|z_1,Z_2}(0|d_s^*, Z_2)$, habiéndose aplicado el Supuesto 1 ($l_s = 0$) para la obtención de la última igualdad.

La formulación (15) es la expresión analítica de la matriz F bajo funciones de pérdida discretas que verifiquen los Supuestos 1-4. Las expresiones (13) y (15) constituyen los resultados correspondientes al objetivo a) que se propuso en la Introducción.

A partir de ahora, el resto del apartado se dedica a resolver lo que en dicha Introducción se denotó como objetivo b): caracterizar situaciones en las que $F = 0$.

[Lema 2]

Bajo las condiciones exigidas en el Lema 1, si el error de previsión u y el vector de regresores $Z = (z_1, Z_2')'$ son estocásticamente independientes, el componente $F_{i,w}$ de la matriz F de McCracken (2000) toma la expresión (13) con:

$$\begin{aligned} R(Z_2) &= H_s Q_s + c \sum_{k \in [1,n-1],\ k \neq s} H_k(1 - 2G_u(0)), \\ Q_s &= c_s + (G_u(l_{s-1}) + G_u(l_{s+1}))(c'_s - c_s) - 2c'_s G_u(0), \end{aligned} \tag{16}$$

siendo G_u la función de distribución incondicional del error de previsión u.

Demostración: Basta aplicar la definición de independencia estocástica en la expresión (15).

[Lema 3]

Bajo las condiciones exigidas en el Lema 2, si, además, G_u es simétrica respecto de $u = 0$, el componente $F_{i,w}$ de la matriz F de McCracken (2000) toma la expresión (13) con:

$$\begin{aligned} R(Z_2) &= H_s Q_s \\ Q_s &= (G_u(l_{s-1}) + G_u(l_{s+1}) - 1)(c'_s - c_s). \end{aligned} \tag{17}$$

Demostración: Por simetría de G_u respecto de $u = 0$, se tiene que $G_u(0) = 1/2 \Rightarrow Q_k = 0$, $k \neq s$, con lo que $R(Z_2) = H_s Q_s$. Además, $Q_s = (G_u(l_{s-1}) + G_u(l_{s+1}))(c'_s - c_s) + c_s - c'_s = (G_u(l_{s-1}) + G_u(l_{s+1}) - 1)(c'_s - c_s)$.

[Corolario 1]

"Sean n modelos de previsión lineales no anidados, con K regresores estacionarios en sentido débil,[16] sea el error de previsión del modelo i-ésimo independiente estocásticamente de los regresores utilizados en dicho modelo y sea la función de distribución para u_i simétrica respecto de cero. Si la función de pérdida f es discreta y verifica los Supuestos 1-4 del apartado 3.3, la fila i-ésima de la matriz F de McCracken (2000) es $F_i = 0_{1\times K}$".

Demostración:

Por el Lema 3 se está verificando (17). Aplicando el Supuesto 2, $Q_s = (G_u(l_{s-1}) + G_u(-l_{s-1}) - 1)(c'_s - c_s)$. Por simetría de G_u, $G_u(x) = 1 - G_u(-x)$ y, en consecuencia, $Q_s = (1 - G_u(-l_{s-1}) + G_u(-l_{s-1}) - 1)(c'_s - c_s) = 0$. Por tanto, $R(Z_2) = 0 \Rightarrow$ aplicando (13)$\Rightarrow F_{i,w} = -E[R(Z_2)] = 0$, para $w = 1, ..., K$.

Éste es el resultado fundamental de la sección. Si todos los modelos cuyas previsiones se evalúan o comparan en el test verifican las condiciones solicitadas en el Corolario 1, todas las filas de F son nulas y la matriz F de McCracken (2000) es una matriz de ceros. Por lo tanto, en dicha situación, los contrastes de habilidad predictiva de un modelo (por ejemplo, C1) o de comparación entre la capacidad de previsión de varios (por ejemplo, DM y Mult2-aprx) verifican la propiedad de irrelevancia asintótica de la incertidumbre paramétrica (IAIP). Este resultado ya fue obtenido por McCracken (2000) para el AE en el mismo contexto, solo que exigiendo únicamente incorrelación donde nosotros exigimos independencia. Por su parte, el SE garantiza $F = 0$ incluso aunque los errores sean asimétricos, y bajo un supuesto aún más débil que la incorrelación, $E(u_{t+1}|Z_{t+1}) = 0$. Sin embargo, la propiedad $F = 0$ no es generalizable a funciones de pérdida no estándar, por lo que el resultado es interesante. Además, nuestra función no es solamente no diferenciable sino ni siquiera continua, y, aún más, representa toda una clase de funciones, y no a una sola función.

A continuación se exponen dos corolarios más, de menor interés, donde se obtiene el mismo resultado anterior, sustituyendo uno de los supuestos utilizados hasta ahora por otro nuevo.

[Corolario 2]

"Sean las mismas condiciones especificadas en Corolario 1, pero sin exigir el Supuesto 2 en la función de pérdidas. Si $K_i = 1$, ie, solo existe un regresor z_1 en el modelo,[17] se verifica también $F_i = 0_{1\times K}$".

Demostración:

Por no existir Z_2, $d_k = \frac{l_k}{\beta_1} \Rightarrow d^*_k = \frac{l_k}{\beta^*_1}$, para cualquier $k = 0, ..., n$. Si $k = s$, $d^*_s = 0$, ya que $l_s = 0$ por Supuesto 1. Por lo tanto, $H_s = g_{z_1}(d^*_s)\frac{d^*_s}{\beta^*_1} = 0$. Dado que siguen cumpliéndose los supuestos de Lema 3, aplicando (17), se tiene $F_{i,w} = R(Z_2) = R = 0$.

[Corolario 3]

"Sean las mismas condiciones especificadas en Corolario 1, pero sin exigir simetría en G_u. Si $K_i = 1$ y g_{z_1} es simétrica respecto de $z_1 = 0$, se verifica también $F_i = 0_{1\times K}$".

Demostración:

$d^*_s = 0 \Rightarrow H_s = 0$, ya que $K_i = 1$, igual que en Corolario 2. Por otro lado, se cumple el Lema 2, luego la expresión (16) queda $R(Z_2) = \underset{k\in[1,n-1],\ k\neq s}{c\sum} H_k(1 - 2G_u(0))$. Como no existe Z_2, $F_{i,w} = R(Z_2) = R$.

Aplicando el Supuesto 2 ($l_k = l_{n-k}$, $k = 1, ..., s$) y la propiedad de simetría en g_{z_1} ($g_{z_1}(x) = g_{z_1}(-x)$), $H_{n-k} = -H_k$, $k = 1, ..., s$. De este modo, $F_{i,w} = c\sum_{k=1}^{s-1} H_k(1 - 2G_u(0)) + c\sum_{k=s+1}^{n-1} H_k(1 - 2G_u(0)) = c\left(\sum_{k=1}^{s-1} H_k - \sum_{k=1}^{s-1} H_k\right)(1 - 2G_u(0)) = 0$.

[16] En el modelo i-ésimo se incluyen $K_i \leq K$ regresores.

[17] Esto excluye también la existencia de constante en el modelo.

3.5. Efecto del incumplimiento de las condiciones para IAIP. Estimación del tamaño asintótico del contraste Diebold-Mariano

En el Corolario 1 se demostró que, bajo ciertos supuestos, el uso de la función de pérdida discreta en los contrastes de capacidad predictiva a los que nos estamos refiriendo, garantiza que la IP no distorsiona las propiedades estadísticas asintóticas de éstos, en el contexto de modelos de previsión lineales, no anidados y con regresores estacionarios.

Las condiciones bajo las que nuestro resultado aplica son: los Supuestos 1-4 respecto a la función de pérdida, simetría respecto de cero en la distribución del error de previsión u_t, e independencia entre error de previsión u_t y regresores Z_t. Los Supuestos 1 y 2 sobre las pérdidas son poco restrictivos. En cambio, las otras dos premisas sobre el tipo de función de pérdida discreta, junto con la hipótesis de simetría en u_t y la de independencia entre u_t y Z_t son dignas de mayor estudio. Por ello, en este apartado vamos a tratar de ilustrar el efecto que tiene su incumplimiento sobre los tests, en términos de su tamaño. Como el Supuesto 4 es solo un caso particular de la idea del Supuesto 3, el impacto sobre F de que no se verificara el primero será del mismo tipo al que se producirá si no se verifica el último. Por este motivo, centraremos nuestras pruebas exclusivamente en el Supuesto 3, para evitar redundancias, y las conclusiones obtenidas serán extensivas al Supuesto 4.

Comprobar el efecto de violar el Supuesto 3 de la función de pérdida discreta o/y la simetría de la distribución de los errores significa simplemente verificar si el resultado expuesto en el Corolario 1 del apartado previo es robusto a dichos supuestos. Esto completaría el trabajo relacionado con el objetivo b) de la Introducción. En cambio, sabemos que el incumplimiento de la condición de independencia entre u_t y Z_t invalida el resultado $F = 0$, igual que pasa bajo SE, impidiendo la IAIP. Lo que pretendemos es cuantificar en qué medida distorsiona las propiedades de los tests bajo f discreta, comparativamente con las funciones SE y AE. Es decir, se trata de solventar la pregunta c) de la Introducción. Entiéndase, por tanto, que se van a responder dos pretensiones de distinta naturaleza.

Para la primera, manteniendo el supuesto de independencia entre u_t y Z_t, vamos a estimar el sesgo asintótico en el tamaño del test DM en un ejercicio de simulación apropiado, bajo dos diseños alternativos. Usaremos en ambos la misma función de pérdida, que incumplirá el Supuesto 3, pero en cada uno se supondrá una distribución de probabilidad para u_t diferente, una simétrica y la otra, asimétrica. Así, no solo tratamos la cuestión de la robustez del Corolario al incumplimiento de los supuestos sobre la función de pérdida discreta, sino que lo haremos en combinación con el incumplimiento de la hipótesis de simetría en u_t. Los resultados sugieren que existe tal robustez, aunque este resultado lo matizaremos después.

Respecto al segundo asunto, vamos a probar qué ocurriría si no tenemos en cuenta la IP a la hora de aplicar el test DM (es decir, usamos incorrectamente S_{ff} en vez de Ω) cuando u_t y Z_t estaban no eran independientes. Fijémonos que la posibilidad de que, en la práctica, aparezca este tipo de relación no es, en absoluto, remota. Por ejemplo, aparecerá siempre que se produzca un error de especificación en el modelo de previsión debido a la omisión de alguna variable explicativa z_j (quizá inobservable) que participaba en el PGD y que estaba correlacionada con otra z_k que sí se ha incluído en el modelo especificado.[18]

Pues bien, estimaremos el sesgo asintótico que se genera en el tamaño del test en estas condiciones, implementándolo con tres funciones de pérdida alternativas: SE, AE y una f discreta. Ésta es una prueba muy importante, ya que, hasta ahora, solo hemos demostrado que $F = 0$ con funciones discretas bajo independencia entre u_t y Z_t, circunstancia que ya sabemos que garantiza también ese mismo resultado para las funciones SE y AE. Nuestro experimento arroja evidencia de que el sesgo asintótico en el tamaño de DM por ignorar la IP es considerablemente menor usando f discreta que usando SE o AE, hasta el punto de que en muchos casos el test podría ser aplicado sin corrección alguna (es decir, usando S_{ff}), sin ver afectado excesivamente su funcionamiento. El experimento que llevamos a cabo en este caso es exactamente el mismo que diseñó West (1996) con regresores y errores de previsión correlados y estimando por 2SLS, solo que nosotros aplicaremos las tres funciones de pérdida mencionadas y estimaremos en esta sección el tamaño asintótico, en vez de en muestras finitas (el caso de muestras finitas será objeto de estudio en el apartado 3.6).

Llegados a este punto, queremos hacer una aclaración. El tamaño asintótico no solo tiene un evidente interés conceptual, sino que además puede estimarse sin necesidad de aplicar el test. Esto es, probablemente, obvio para el lector, pero adjuntamos a continuación el razonamiento: el tipo de contrastes a los que se refieren los teoremas de West (1996) y McCracken (2000) son de la forma especificada en (1a), en

[18] En este tipo de situaciones, procede estimar usando variables instrumentales, para evitar la inconsistencia del estimador.

general, o en (8), en particular. En concreto, el que nos interesa, el test DM, es de la forma (8).[19] Sea $d = f_1 - f_2$, y sean $\Omega^{(2)} = \Omega_{11} + \Omega_{22} - 2\Omega_{12}$ la varianza asintótica corregida de la IP para el test DM y $S_{ff}^{(2)} = S_{ff\ 11} + S_{ff\ 22} - 2S_{ff\ 12}$ la varianza sin dicha corrección. Nótese que ésta última es, por ejemplo, la varianza utilizada en el Capítulo 2 de la tesis para la aplicación de este contraste. La especificación correcta del test debería ser $\vartheta = \sqrt{P}\frac{\overline{d}-E(d)}{\sqrt{\Omega^{(2)}}} \overset{a}{\sim} N(0,1)$, de modo que no se rechazaría $H_0 \equiv E(d) = 0$ si $\lambda_{\alpha/2} \leq \vartheta \leq \lambda_{1-\alpha/2}$, siendo $\lambda_{\alpha/2}$ y $\lambda_{1-\alpha/2}$ los percentiles $\alpha/2$ y $1-\alpha/2$ de una distribución $N(0,1)$. Por tanto, si se emplea erróneamente la varianza $S_{ff}^{(2)}$ en vez de $\Omega^{(2)}$, se rechaza H_0 si el estadístico $S_1 = \vartheta\sqrt{\frac{\Omega^{(2)}}{S_{ff}^{(2)}}}$ verifica $\lambda_{\alpha/2} \leq S_1 \leq \lambda_{1-\alpha/2}$. Dado que la distribución asintótica de S_1 es conocida, en concreto, $S_1 \sim (0, \frac{\Omega^{(2)}}{S_{ff}^{(2)}})$, entonces el sesgo asintótico incurrido por el test DM al aplicarse con varianza $S_{ff}^{(2)}$, es decir, sin incorporar el efecto de la estimación paramétrica, es:

$$sesgo\ DM = 100(P^* - \alpha), \tag{18}$$
$$P^* = 1 - P\left[\lambda_{\alpha/2} \leq S_1 \leq \lambda_{1-\alpha/2}\right], \text{ siendo } S_1 \text{ una } v.a.\ N(0, \frac{\Omega^{(2)}}{S_{ff}^{(2)}}),$$

donde P^* simboliza la probabilidad de rechazar H_0 al aplicar el test DM con varianza $S_{ff}^{(2)}$.

Por consiguiente, la estimación del sesgo asintótico de DM se reduce a estimar las varianzas $S_{ff}^{(2)}$ y $\Omega^{(2)}$ y aplicar luego la expresión (18).

Por cuestiones de simplicidad en la presentación de los diseños de los experimentos, invertimos el orden natural de su exposición: primero, el ejercicio correspondiente a la correlación entre variables explicativas y errores de previsión, y después, el relacionado con la violación de supuestos sobre la especificación particular de la función de pérdidas discreta y sobre la simetría de la distribución de los errores.

3.5.1. Errores de previsión correlados con los regresores (Experimento A)

Diseño Experimento Realizaremos el mismo experimento de simulación que West (1996) para errores de previsión correlacionados con los regresores del modelo, para tres funciones de pérdida: una discreta (que se detallará más abajo), SE y AE. Para cada una de ellas, la simulación permitirá estimar $\Omega^{(2)}$ y $S_{ff}^{(2)}$, a partir de las expresiones (1b) a (5), y, aplicando (18), se tendrán calculados los sesgos asintóticos del test DM en cada caso. Los resultados teóricos asintóticos de West (1996) y McCracken (2000) dependen del parámetro π (véase apartado 2.1, cuarto supuesto), cuya estimación habitual es el ratio P/R. Pues bien, utilizaremos los mismos casos empleados en West (1996) para dicho ratio: $P/R = 0{,}25$, 0,50, 1, 2, 3, 4, 6 y 7, añadiendo nosotros el caso $P/R = 0{,}1$. Sin embargo, nosotros fijaremos $T = P + R - 1 = 1000$,[20] en vez de usar el valor de West (1996), $T = 200$, precisamente porque buscamos el cálculo del tamaño asintótico del test y no del tamaño en muestras finitas, por lo que requerimos P suficientemente grande. De este modo, se llevará a cabo un experimento para cada uno de los siguientes pares R, P: (a) $R = 910$, $P = 101$ ($\pi = 0{,}1$); (b) $R = 801$, $P = 200$ ($\pi = 0{,}25$); (c) $R = 667$, $P = 334$ ($\pi = 0{,}50$); (d) $R = 501$, $P = 500$ ($\pi = 1$); (e) $R = 334$, $P = 667$ ($\pi = 2$); (f) $R = 250$, $P = 751$ ($\pi = 3$); (g) $R = 200$, $P = 801$ ($\pi = 4$); (h) $R = 143$, $P = 858$ ($\pi = 6$); (i) $R = 125$, $P = 876$ ($\pi = 7$).

El resto del diseño del experimento es exactamente el propuesto en la sección 5.2 de West (1996):

[PGD] $y_t = w_{1t} + w_{2t} + v_t$, siendo $w_{it} = z_{it} + dv_t$, $d = 1$, $(z_{1t}, z_{2t}, v_t)' \overset{iid}{\sim} N(0_{3\times1}, I_{3\times3})$.
[M_1] $y_t = \beta_{10} + \beta_{11}w_{1t} + u_{1t} = X_{1t}'\beta_1 + u_{1t}$, con $\beta_1 = (\beta_{10}, \beta_{11})'$ y $X_{1t} = (1, w_{1t})'$.
[M_2] $y_t = \beta_{20} + \beta_{22}w_{2t} + u_{2t} = X_{2t}'\beta_2 + u_{2t}$, con $\beta_2 = (\beta_{20}, \beta_{22})'$ y $X_{2t} = (1, w_{2t})'$.

Por lo tanto, $\beta = (\beta_1', \beta_2')'$ y $\beta_{10}^* = \beta_{20}^* = 0$, $\beta_{11}^* = \beta_{22}^* = 1$ ($\beta^* = (0,1,0,1)'$). Como se pretendía, los regresores $w_{it} = z_{it} + v_t$ y los errores de previsión $u_{it}(\beta^*) = w_{jt} + v_t$ ($i = 1$, $j = 2$ e $i = 2$, $j = 1$)

[19] El razonamiento a continuación se llevaría a cabo de forma análoga para los tests más generales especificados por (1a).
[20] La muestra disponible es de tamaño $T + 1 = P + R$. Repásese la introducción a la sección 2.

están correlados,[21] con coeficiente de correlación $\rho_{u_i w_i} = \frac{2\sigma_v^2}{(\sigma_{z_2}^2+2\sigma_v^2)^{1/2}(\sigma_{z_1}^2+\sigma_v^2)^{1/2}} = \sqrt{\frac{2}{5}}$. Por otro lado, no es difícil demostrar que se cumple $E(f_{1t}) = E(f_{2t})$ tanto para f discreta como cuadrática,[22] así que la hipótesis nula del test DM (igual capacidad predictiva de M_1 y M_2) es cierta, y el experimento corresponde a un ejercicio de tamaño. Respecto a la estimación de los modelos y a la previsión, basta reseñar que la estimación de los parámetros β_{i0} y β_{ii} se llevará a cabo por 2SLS, para evitar la inconsistencia del estimador, empleando los instrumentos $A_{it} = (1, z_{it})'$. Es decir, $\widehat{\beta}_{it} = \left(t^{-1}\sum_{s=1}^{t} A_{is}X_{is}'\right)^{-1}\left(t^{-1}\sum_{s=1}^{t} A_{is}y_s\right)$.[23] El horizonte de previsión es uno, por lo que el dato y_{t+1} se prevé con $v_{i,t+1} = X_{it+1}'\widehat{\beta}_{it}$. El error de previsión es $u_{it+1} = y_{t+1} - X_{it+1}'\beta_i^*$ y su estimación es $\widehat{u}_{it+1} = y_{t+1} - v_{i,t+1}$.

Las funciones de pérdida SE y AE son bien conocidas (u_{it+1}^2 y $|u_{t+1}|$, respectivamente). La función de pérdida discreta que usamos es la definida por la partición P_A y por la matriz (9), siendo $P_A = \{l_0 = -\infty, l_1 = -0{,}75\sigma_y, l_2 = 0, l_3 = +0{,}75\sigma_y, l_4 = +\infty\}$, y denotando por σ_y la desviación típica de y_t.Véase que dicha función está verificando los Supuestos 1-4 del apartado 3.3.

Cada ejercicio (definido por una combinación P, R) se repite 5000 veces, generándose, por tanto, 5000 estimaciones de las matrices $S_{ff}^{(2)}$ y $\Omega^{(2)}$. Los detalles de la estimación en cada repetición se exponen en el punto siguiente. Finalmente, la media muestral de esas 5000 matrices obtenidas constituirá la estimación definitiva de $S_{ff}^{(2)}$ y $\Omega^{(2)}$ que deberá utilizarse en la expresión (18) para determinar los sesgos asintóticos.

Respecto a la estimación de las varianzas $S_{ff}^{(2)}$ y $\Omega^{(2)}$ en cada repetición debe comentarse lo siguiente:

a) Estimación de las matrices S_{ff}, S_{fh}, S_{hh} y B:

Se estiman usando las expresiones en (2)−(4), pero incluyendo solo el retardo $j = 0$, puesto que no hay autocorrelación en f_t ni h_t ni correlación cruzada no contemporánea entre ambas, y sustituyendo las covarianzas poblacionales por las muestrales. Es decir, $\widehat{S}_{ff} = P^{-1}\sum_{t=R}^{T}\left(f_{t+1}(\widehat{\beta}_t) - \overline{f}\right)\left(f_{t+1}(\widehat{\beta}_t) - \overline{f}\right)'$, $\widehat{S}_{fh} = P^{-1}\sum_{t=R}^{T}\left(f_{t+1}(\widehat{\beta}_t) - \overline{f}\right)h_t'$ y $\widehat{S}_{hh} = T^{-1}\sum_{t=1}^{T} h_t h_t'$, siendo $h_t = (A_{1t}'\widetilde{u}_{1t},\ A_{2t}'\widetilde{u}_{2t})'$, con $\widetilde{u}_{it} = y_t - X_{it}'\widehat{\beta}_{iT}$. La expresión para h_t se deduce fácilmente de la expresión para $\widehat{\beta}_t - \beta^*$ en este ejercicio y del enunciado del Supuesto 2 del apartado 2.1.1, correspondiente al teorema de West (1996).

Respecto a las realizaciones $f_{t+1}(\widehat{\beta}_t)$ de la función de pérdidas, se tendrá $f_{t+1}(\widehat{\beta}_t) = (\widehat{u}_{1t+1}^2, \widehat{u}_{2t+1}^2)'$ si f es el SE, y $f_{t+1}(\widehat{\beta}_t) = (|\widehat{u}_{1t+1}|, |\widehat{u}_{2t+1}|)'$ si es el AE, con $\widehat{u}_{it+1} = y_{t+1} - X_{it+1}'\widehat{\beta}_{it}$. Si f se refiere a la función discreta, $f_{it+1}(\widehat{\beta}_{it}) = a_{jk}$, si $l_{j-1} \le y_{t+1} \le l_j$, $l_{k-1} \le v_{i,t+1} = X_{it+1}'\widehat{\beta}_{it} \le l_k$, siendo los valores l_r y a_{jk} los especificados arriba en la definición de la partición P_A y matriz (9) que definen, a su vez, la función de pérdidas discreta que empleamos.

Por su parte, la matriz B se estima por $\widehat{B}_T = \begin{pmatrix} \left(T^{-1}\sum_{s=1}^{T} A_{1s}X_{1s}'\right)^{-1} & 0_{2\times 2} \\ 0_{2\times 2} & \left(T^{-1}\sum_{s=1}^{T} A_{2s}X_{2s}'\right)^{-1} \end{pmatrix}$, de nuevo en virtud del Supuesto 2 del apartado 2.1.1 y de la expresión para $\widehat{\beta}_t - \beta^*$ en este ejercicio. Por su parte, Π se estimará según (5), con $\widehat{\pi} = P/R$.

[21] La correlación entre errores de previsión y regresores aparece en la expresión de la matriz F (véase, por ejemplo, el apartado 2.1), matriz que está evaluada en $\beta = \beta^*$. Por eso, la definición relevante del error de previsión es $u_t(\beta^*)$.

[22] En el caso cuadrático, la hipótesis se reduce a $E(u_{1t}^2) = E(u_{2t}^2)$, y la demostración es inmediata. En el caso discreto, es también sencilla, aunque algo más trabajosa, debiendo demostrarse, $\forall j,k$, que la probabilidad $P(l_{j-1} \le y_{t+1} \le l_j,\ l_{k-1} \le v_{i,t+1} \le l_k)$ es igual para $i = 1$ e $i = 2$.

[23] Como es bien sabido, el estimador 2SLS utiliza como variable instrumental asociada a cada regresor del modelo la mejor combinación lineal de los instrumentos disponibles. En este caso, dado que los únicos instrumentos que se disponen para $X_{it} = (1, w_{it})'$ son $A_{it} = (1, z_{it})'$, es obvio que las variables instrumentales óptimas son precisamente 1 (para el regresor 1) y z_{it} (para w_{it}).

b) Estimación de la matriz F:

La matriz F de dimensión 2×4 contendrá ceros en todas las posiciones salvo $F_{1,2} = \left[\frac{\partial E f_1}{\partial \beta_{11}}\right]_{\beta=\beta^*}$ y $F_{2,4} = \left[\frac{\partial E f(\beta)}{\partial \beta_{21}}\right]_{\beta=\beta^*}$.[24] Tal y como obtuvo West (1996), en el caso del SE, las expresiones de dichas derivadas son $F_{1,2} = -2E(u_{1t}w_{1t})$ y $F_{2,4} = -2E(u_{2t}w_{2t})$. Por su parte, McCracken (2000) demuestra que, si f es el AE, entonces $F_{1,2} = -E\left[sign(u_{1t})w_{1t}\right]$ y $F_{2,4} = -E\left[sign(u_{2t})w_{2t}\right]$,[25] (véase apartado 2.1). Por lo tanto, las estimaciones de estos elementos de la matriz F para estas dos funciones de pérdida son, obviamente, $\widehat{F}_{1,2} = -2P^{-1}\sum_{t=R}^{T}\widehat{u}_{1t+1}w_{1t+1}$ (en el caso del SE) y $\widehat{F}_{1,2} = -P^{-1}\sum_{t=R}^{T} sign(\widehat{u}_{1t+1})w_{1t+1}$ (en el caso del AE), siendo análoga la estimación $\widehat{F}_{2,4}$.

Por su parte, la estimación de $F_{1,2}$ y $F_{2,4}$ en el caso de f discreta se realiza siguiendo la expresión (13). En este caso, cada modelo M_i solo emplea un regresor y una constante, por lo que las variables de dicha expresión son $z_1 = w_i$ y $Z_2 = 1$ (constante). Con ello, en este caso, $F_{1,2}$ queda reducido a $F_{1,2} = -\sum_{k=1}^{3} H_k Q_k$, siendo $H_k = g_{w_1}(d_k^*)\frac{d_k^*}{\beta_{11}^*}$, $d_k^* = \frac{l_k - \beta_{10}^*}{\beta_{11}^*}$, $Q_k = \sum_{j=1}^{4} B(j,k)(a_{jk} - a_{jk+1})$, $B(j,k) = G_{u_1|w_1}(l_j - l_k|d_k^*) - G_{u_1|w_1}(l_{j-1} - l_k|d_k^*)$ (son las definiciones del apartado 3.2, solo que teniendo en cuenta que $Z_2 = 1$ y no es, por tanto, variable aleatoria). Dado el diseño del ejercicio, g_{w_1} se corresponderá con la función de densidad de una distribución $N(0,1)$ y $G_{u_1|w_1}$ con la función de distribución Normal de u_1 condicionada a w_1, cuya expresión coincide, como es bien sabido, con la de $N(b(w_1), \sigma_{u_1}^2(1-\rho_{u_1w_1}^2))$, siendo $b(w_1) = \mu_{u_1} + (w_1 - \mu_{w_1})\rho_{u_1w_1}\frac{\sigma_{u_1}}{\sigma_{w_1}}$. En este caso, la distribución resultante es $N(w_1, 3)$, dado que $\mu_{u_1} = \mu_{w_1} = 0$, $\sigma_{u_1} = \sqrt{5}$, $\sigma_{w_1} = \sqrt{2}$ y $\rho_{u_1w_1} = \sqrt{\frac{2}{5}}$. Para estimar $F_{1,2}$, se utilizarán estas dos distribuciones, junto con las expresiones anteriores de H_k y Q_k, sustituyendo β_1^* por su estimación $\widehat{\beta}_{1T}$. La estimación de $F_{2,4}$ se realiza de forma análoga.

Fijémonos que la estimación de F en el caso de f discreta es más compleja que si usamos SE ó AE, sobre todo por la necesidad de conocer las distribuciones $g_{z_1|Z_2}$ y $G_{u|z_1,Z_2}$, problema que ya fue expuesto por McCracken (2004), y que afecta a casi cualquier función de pérdidas no diferenciable. Sin embargo, el empleo de f discreta seguirá siendo ventajoso, porque, tal y como probaremos con este experimento de simulación, el efecto sobre el sesgo asintótico del test DM cuando se emplea $S_{ff}^{(2)}$ en vez de $\Omega^{(2)}$ es poco relevante incluso en casos de correlación no nula entre regresores y errores de previsión, siempre que se utilice precisamente dicha función de pérdida, por lo que, si los tamaños muestrales son elevados, el usuario puede ahorrarse no solo la estimación de F sino también la de las matrices S_{fh}, S_{hh}, y B, y bastará con estimar S_{ff}.

En este punto, es necesario hacer una aclaración. Si, dada una situación de dependencia entre error de previsión y regresores, el usuario desea en cualquier caso estimar F habiendo utilizado f discreta (por ejemplo, porque se trate de una aplicación donde sea crucial garantizar la precisión del contraste), o bien se realizan supuestos sobre las distribuciones $g_{z_1|Z_2}$ y $G_{u|z_1,Z_2}$ y se aplica la expresión (15),[26] o bien se emplea el mecanismo de estimación presentado por McCracken (2004), que no requiere tales supuestos.

Resultados En la Tabla 1 se exponen las estimaciones obtenidas respecto al sesgo asintótico en tamaño para el test DM cuando la varianza considerada en el contraste no se corrige de IP. En ella puede verse que, como ya adelantamos arriba, el uso de la función de pérdida discreta garantiza que el sesgo asintótico en el tamaño del test DM empleando $S_{ff}^{(2)}$ en la aplicación del contraste, en vez de $\Omega^{(2)}$, sea razonablemente pequeño. Para un tamaño teórico del 5 %, el sesgo asintótico con f discreta se situó en todos los casos entre 0,4 % y 6 %, mientras con el SE, lo hizo entre 4 % y 38 %, y algo similar con el AE. En la práctica, las situaciones predictivas más habituales son las que corresponden a $\pi \leq 1$. Para éstas, el sesgo bajo f

[24] Esta forma de definir F es válida también para el SE, ya que la continuidad de f garantiza que $F = E\left(\left[\frac{\partial f(\beta)}{\partial \beta}\right]_{\beta=\beta^*}\right) = \left[\frac{\partial E f(\beta)}{\partial \beta}\right]_{\beta=\beta^*}$.

[25] $sign(x)$ es una función que toma valor 1 si $x \geq 0$ y -1 si $x < 0$.

[26] Ésa es la expresión correcta siempre que la función discreta diseñada verifique los Supuestos 1 a 4 del apartado 3.3. En caso contrario, usar la expresión general (13).

discreta siempre es inferior a 2,6 % (el tamaño alcanzaría 7,6 % en vez del 5 % teórico), frente al 23 % del SE o el 19 % del AE (tamaños de 28 % y 24 %, respectivamente, cuando el teórico era 5 %).

Por supuesto, se confirma que el sesgo del contraste se reduce al disminuir $\widehat{\pi}$, para las tres funciones de pérdida examinadas, tal y como afirma la teoría al respecto.[27]

TABLA 1. Sesgo asintótico Tamaño DM. Experimento A

			Sesgo asintótico (%). $\alpha = 5\,\%$		
$\widehat{\pi} = P/R$	R	P	f discreta	f = SE	f = AE
0,10	910	101	0,40	4,20	3,33
0,25	801	200	0,92	9,40	7,44
0,50	667	334	1,62	15,56	12,50
1	501	500	2,66	22,88	18,86
2	334	667	3,93	29,58	24,99
3	250	751	4,69	32,68	27,93
4	200	801	5,23	34,77	29,93
6	143	858	5,93	37,32	32,42
7	125	876	6,15	37,95	33,02

SE (AE): cuadrado (valor absoluto) del error de previsión.

La razón de este resultado favorable a f discreta en comparación con SE y AE se encuentra en el valor que toma la matriz F. Para comprobarlo, se presentan en el Cuadro 1 las estimaciones de los distintos términos que componen la varianza correcta bajo IP en el test DM, $\Omega^{(2)}$, para un ejercicio concreto, el correspondiente al caso $\widehat{\pi} = 1$. Utilizando la expresión (1b), Ω se puede escribir como $\Omega = S_{ff} + T_1 + T_2$, siendo $T_1 = \Pi(FBS'_{fh} + S_{fh}B'F')$ y $T_2 = 2\Pi F V_\beta F'$. Así mismo, podemos escribir $\Omega^{(2)} = S^{(2)}_{ff} + T^{(2)}_1 + T^{(2)}_2$, donde $T^{(2)}_i$ se define de forma análoga a cómo se hizo para $S^{(2)}_{ff}$ y $\Omega^{(2)}$ en el apartado 2.2. El efecto clave de F para que el sesgo en tamaño para f discreta sea menor que para las otras dos funciones se produce en el término $T^{(2)}_2$, que incluye dos multiplicaciones por la matriz F en su expresión, por una sola en el término $T^{(2)}_1$. Por ejemplo, la matriz F bajo AE es aproximadamente 2,85 veces la matriz F bajo f discreta, por lo que el término $T^{(2)}_2$ es más de ocho veces el análogo para f discreta. Así, en el caso de AE, se tiene que $T^{(2)}_2 = 1,74 S^{(2)}_{ff}$, mientras con f discreta, $T^{(2)}_2 = 0,23 S^{(2)}_{ff}$. Como el término $T^{(2)}_1$ resulta prácticamente nulo en ambos casos, el resultado final es que, usando el AE, la estimación paramétrica añade una varianza $T^{(2)}_1 + T^{(2)}_2$ de magnitud igual a 1,77 veces la varianza $S^{(2)}_{ff}$, mientras la varianza añadida usando f discreta solo es 0,22 veces $S^{(2)}_{ff}$. Las diferencias con el caso del SE aún son más ostensibles.

CUADRO 1. Estimación varianza del estadístico DM, bajo IP. Experimento A. Caso $\widehat{\pi} = 1$

f	Matriz F	$S^{(2)}_{ff}$	$T^{(2)}_1$	$T^{(2)}_2$	$\Omega^{(2)}$	$\frac{T^{(2)}_1}{S^{(2)}_{ff}}$	$\frac{T^{(2)}_2}{S^{(2)}_{ff}}$	$\frac{T^{(2)}_1+T^{(2)}_2}{S^{(2)}_{ff}}$	$\frac{\Omega^{(2)}}{S^{(2)}_{ff}}$
discreta	$\begin{pmatrix} 0 & -0,25 & 0 & 0 \\ 0 & 0 & 0 & -0,25 \end{pmatrix}$	1,36	0,00	0,31	1,67	0,00	0,23	0,22	1,22
SE	$\begin{pmatrix} 0 & -4,01 & 0 & 0 \\ 0 & 0 & 0 & -4,02 \end{pmatrix}$	36,31	2,31	80,31	118,93	0,06	2,21	2,28	3,28
AE	$\begin{pmatrix} 0 & 0,71 & 0 & 0 \\ 0 & 0 & 0 & 0,71 \end{pmatrix}$	1,45	0,04	2,53	4,02	0,02	1,74	1,77	2,77

SE (AE): cuadrado (valor absoluto) del error de previsión.

[27] Véase la expresión de Ω en (1b) y la relación creciente entre Π y π en (5). De hecho, si $\pi = 0$, entonces $\Omega = S_{ff}$ y, por tanto, la IP sería irrelevante.

3.5.2. Incumplimiento de supuestos sobre la estructura de la función de pérdidas discreta (Experimento B)

El objetivo que nos proponemos ahora es comprobar el efecto que, sobre los resultados demostrados en nuestro Corolario 1, pudiera derivarse del incumplimiento del Supuesto 3 respecto a la función de pérdidas discreta y de la simetría en los errores, que son los requisitos más restrictivos. Para ello, llevaremos a cabo el mismo experimento de simulación anterior (Experimento A), solo que sin generar correlación entre regresores y errores de previsión, es decir, con $d = 0$. Ahora solo incluiremos funciones de pérdida discretas. Las función elegida será (19), que incumple el Supuesto 3. Se usará tanto bajo un contexto de simetría en u_t como de asimetría (pero verificándose $E(u_t) = 0$). Por lo tanto, se tendrán dos experimentos de simulación, que tratan de estimar el impacto de la violación del Supuesto 3, tanto cuando el resto de supuestos se cumplen, como cuando, además, falla la simetría en la distribución de u_t.

El diseño del experimento, por tanto, es el mismo del Experimento A, pero con $d = 0$. Es decir, cada uno de los modelos de previsión M_i utiliza un regresor distinto $w_i = z_i \sim N(0, 1)$ pero omite otro de los regresores que realmente forman parte del PGD. Vuelve a ser un ejercicio de tamaño, ya que $E(u_{1t}^2) = E(u_{2t}^2)$ y la hipótesis nula de igualdad de capacidad predictiva entre M_1 y M_2 es cierta. Como z_{it} y u_{it} son independientes estocásticamente, la estimación de los parámetros se realiza ahora por el método OLS.

Respecto a la estimación de las matrices S_{ff}, S_{hh}, S_{fh}, B involucradas en Ω, es válido todo lo comentado arriba para la función discreta, solo que ahora $A_{is} = X_{is}$. Por su parte, la estimación de la matriz F se simplifica, al ser en este caso las distribuciones a considerar todas incondicionadas. Así, $F_{1,2}$ y $F_{2,4}$ se estiman de la manera que se describió en el Experimento A respecto a funciones discretas pero siendo ahora $B(j, k) = G_{u_i}(l_j - l_k) - G_{u_i}(l_{j-1} - l_k)$, donde G_{u_i} se corresponde con la función de distribución del error de previsión u_i (que puede ser simétrica o asimétrica, según de cuál de los dos ejercicios se trate).

La función de pérdida considerada responde a la misma partición P_A especificada en el Experimento A, dividiendo la recta real en 4 regiones. Las asignaciones de pérdidas de las funciones que incumplen el Supuesto 3 son:

		$v_{i,t}$			
		G−	P−	P+	G+
y_t	G−	0	1	2	15
	P−	1	0	2	10
	P+	3	2	0	1
	G+	3	2	1	0

(19)

Véase que, efectivamente, (19) no verifica el Supuesto 3 del apartado 3.3, debido al valor de los cuadrantes (1,4) y (2,4). Si denotamos por a_{jk} la pérdida asignada al cuadrante (j,k), el Supuesto 3 exigía que, para todo $k \neq s = 2$,[28] se verificaran las condiciones $a_{jk+1} - a_{jk} = +c$ si $k \geq j$ y $a_{jk+1} - a_{jk} = -c$ si $k < j$. Dicha condición se cumple en todos los casos (con $c = 1$) salvo en $j = 1$, $k = 4$, y $j = 2$, $k = 4$, ya que $a_{14} - a_{13} = +13$ y $a_{24} - a_{23} = +8$.

Finalmente, queda por especificar cómo se introducirá la asimetría en u_t en el experimento que la requiere. Se generarán 200000 realizaciones de una distribución $CN(e, \mu_1, \mu_2, \sigma_1, \sigma_2)$, y se obtendrán estimaciones de los puntos de la función de distribución asociada que se necesitan para la construcción de la matriz F. La distribución CN es una "mezcla" de distribuciones Normales, cuya función de distribución G_{CN} viene definida por:

$$G_{CN}(x) = (1 - e)\Phi_1(x) + e\Phi_2(x),$$

denotando por $\Phi_i(.)$ la función de distribución de una variable aleatoria $N(\mu_i, \sigma_i)$. Eligiendo los valores paramétricos a continuación, se obtiene una distribución asimétrica pero de media cero, concentrando el 70 % de la probabilidad en puntos $x < 0$, y solo el 30 % en puntos $x > 0$: $e = 0{,}35$, $\mu_1 = -2{,}154$, $\mu_2 = 4$, $\sigma_1 = 1$ y $\sigma_2 = 5$.

[28] Recuérdese que $s = n/2$, siendo n el número de regiones de la partición utilizada en la definición de la función de pérdidas discreta.

En el experimento que se lleva a cabo en un escenario de simetría en u_t, la distribución utilizada será la $N(0,1)$.

Para terminar con las aclaraciones respecto al experimento, queda por comentar que, a diferencia de la implementación del Experimento A, ahora solo realizaremos el ejercicio para los valores de π más extremos ($\pi = 0{,}1$ y $\pi = 7$), simplemente por simplificar el ejercicio.

Los resultados obtenidos se muestran en la Tabla 2, a continuación. Se trata del sesgo asintótico estimado para el tamaño del test DM cuando no se incorpora en el test la IP realmente existente, en un contexto de independencia entre errores de previsión y regresores y con funciones de pérdida discretas. Como sabemos, en esta situación, si se verificara el Supuesto 3, además del resto de condiciones establecidas en Corolario 1, el sesgo es cero, como se probó allí.

TABLA 2. Sesgo asintótico Tamaño DM. Experimento B

			Sesgo asintótico (%). $\alpha = 5\,\%$.	
			Función (19)	
$\widehat{\pi} = P/R$	R	P	Simetría en u_i	Asimetría en u_i
0,10	910	101	0,030	0,038
7	125	876	0,433	0,551

Tal y como se observa en la Tabla 2, el incumplimiento del Supuesto 3 y de la hipótesis de simetría en la distribución de los errores de previsión prácticamente no tuvo ningún coste en términos de tamaño asintótico, en los ejercicios diseñados.[29] Sin embargo, esto no se produjo porque la matriz F se anulara, pero sí resultó suficientemente pequeña como para que, en combinación con los valores del resto de matrices involucradas en $\Omega - S_{ff}$, dicha diferencia fuera muy pequeña y el tamaño muy próximo al correcto. Podemos comprobar esto en el Cuadro 2, en el que se presentan las estimaciones obtenidas para los términos que afectan a la varianza correcta bajo IP para el estadístico DM ($\Omega^{(2)}$), para uno de los ejercicios realizados (en concreto, el correspondiente a distribución de u_i asimétrica y $\widehat{\pi} = 7$).[30] Aunque $F_{1,2}$ y $F_{2,4}$ no son cero, el término $T_1^{(2)}$, asociado a la matriz $\Pi(FBS'_{fh} + S_{fh}B'F')$ de (1b), es prácticamente nulo, y el término $T_2^{(2)}$, asociado a la matriz $2\Pi F V_\beta F'$, es muy pequeño. De este modo, la diferencia entre la varianza correcta del test DM ($\Omega^{(2)}$) y la que no incorpora IP ($S_{ff}^{(2)}$) resultó mínima.

CUADRO 2. Estimación de la varianza del estadístico DM. Experimento B. Caso: Asimetría en u_i y $\widehat{\pi} = 7$

f	Matriz F	$S_{ff}^{(2)}$	$T_1^{(2)}$	$T_2^{(2)}$	$\Omega^{(2)}$	$\frac{T_1^{(2)}}{S_{ff}^{(2)}}$	$\frac{T_2^{(2)}}{S_{ff}^{(2)}}$	$\frac{T_1^{(2)}+T_2^{(2)}}{S_{ff}^{(2)}}$	$\frac{\Omega^{(2)}}{S_{ff}^{(2)}}$
discreta	$\begin{pmatrix} 0 & -0,18 & 0 & 0 \\ 0 & 0 & 0 & -0,18 \end{pmatrix}$	1,98	0,00	0,09	2,07	0,00	0,05	0,05	1,05

La conclusión que se extrae de estos resultados es doble:

a) Por un lado, las simulaciones muestran que *el Corolario 1 no es robusto al incumplimiento de los supuestos* que añaden una estructura particular en la función de pérdida discreta, ya que, al violarse el Supuesto 3 (y análogamente hubiera ocurrido con el Supuesto 4), la matriz F ya no es cero. Esta conclusión es muy relevante, porque da respuesta a una importante cuestión que, implícitamente, se suscita tras derivar el Corolario 1. La función de pérdida discreta describe una clase de funciones de gran amplitud. En el límite, cualquier función de pérdida podría aproximarse por una función discreta genérica como las introducidas en el apartado 2.3, en base a una partición suficientemente fina. Los Supuestos 1 y

[29] Para confirmar estos resultados, se han llevado a cabo experimentos análogos a éstos, pero implementados del modo convencional, es decir, ejecutando el test DM en cada repetición, y estimando el tamaño a través del número de rechazos de la hipótesis nula obtenidos. Aunque el experimento se realiza en un contexto de muestras finitas, se toman valores de R y P suficientemente grandes (en concreto, se usó $R = 10000$ y $P = 1000$), de modo que el tamaño obtenido se puede considerar asintótico, para un valor $\pi = \frac{P}{R}$ (en nuestro caso, $\pi = 0,1$). Los resultados que se lograron están en concordancia con los mostrados en la Tabla 2.

[30] Se presentan los resultados de un solo ejercicio simplemente para evitar redundancias, ya que las conclusiones son las mismas en todos los casos. Se eligió éste en concreto porque es el que generó más sesgo.

2 prácticamente no restringen el conjunto de posibles funciones de pérdida bajo las que el resultado del Corolario 1 aplica. En cambio, los Supuestos 3 y 4 sí lo hacen. Pero si las simulaciones sugirieran que el resultado del corolario fuera realmente robusto a ellos y a la asimetría en la distribución de los errores, podríamos inferir que, en un contexto de modelos lineales no anidados cuyos errores de previsión y variables explicativas fueran independientes, la propiedad de IAIP se tiene para cualquier función de pérdida. Ahora queda claro que la función de pérdida discreta debe tener una estructura particular mínima, la añadida por los Supuestos 3 y 4, para garantizar $F = 0$.

b) Sin embargo, las simulaciones anteriores también muestran que, aunque alguno o varios de los supuestos del Corolario 1 no se cumplan y, por tanto, la matriz F no se anule, el sesgo asintótico en el tamaño de tests como DM puede ser prácticamente cero bajo f discreta (en el contexto de modelos lineales con variables estacionarias con errores y regresores independientes). Esto ocurre porque la función discreta tiende a generar matrices F con valores pequeños, en general, pero necesitará otras condiciones relacionadas con el marco predictivo que garanticen que los valores de S_{fh} y V_β sean adecuados para que los dos sumandos que componen $\Omega - S_{ff}$ sigan siendo pequeños. Es posible que la magnitud de los valores de F en situaciones de incumplimiento de los Supuestos 3 ó 4 no sea independiente de la definición de la partición asociada a la función de pérdida, de modo que, cuanto más cercana a una continua fuera la función discreta, mayor sea el sesgo en tamaño en situaciones en las que las premisas comentadas no se verifican. Por su parte, muchas cuestiones pueden tener influencia en el sesgo de tamaño a través de S_{fh} y V_β, como el número de regresores, el grado de relación entre ellos, su estructura univariante, etc. No obstante, dejamos abierta para futuras investigaciones la delimitación del conjunto de condiciones bajo las que, pese a no cumplirse los Supuestos 3 ó 4 o/y la simetría de los errores, el sesgo asintótico tiende a ser nulo en tests que usan f discreta en contextos de modelos de previsión lineales con variables estacionarias independientes del error de previsión.

3.6. Ejercicios de simulación para muestras finitas

Para terminar con esta sección, estudiaremos de nuevo el funcionamiento del test DM cuando la IP no se corrige, solo que ahora centraremos el análisis en muestras finitas. Hasta ahora solo se han presentado resultados de validez asintótica, así que llevaremos a cabo los experimentos anteriores usando muestras cortas. En tal caso, dado que ya no se dispone de una expresión para el tamaño del test, se realizarán las simulaciones de forma convencional. Es decir, ya no se tratará de estimar $S_{ff}^{(2)}$ y $\Omega^{(2)}$ y aplicar la expresión (18), sino de ejecutar los contrastes en cada repetición del ejercicio y contabilizar el número de rechazos de la hipótesis nula, como en cualquier experimento habitual de estimación del tamaño de un test. Vamos a aplicar el test DM bajo tres tipos de funciones de pérdida f: el SE, el AE y una función de pérdidas discreta, que será, en concreto, la misma que se utilizó en el Experimento A (partición P_A y matriz de pérdidas (9)). Cuando f sea la función discreta, además de la versión habitual de DM, emplearemos también una versión alternativa del test, que denotamos en el Capítulo 2 de la tesis como Mult2-aprx, cuya única diferencia respecto a la implementación estándar radica en la estimación propuesta para la varianza del estadístico de diferencia de medias muestrales que usa DM. Nuevamente, ejecutamos los contrastes ignorando conscientemente el asunto de la IP (es decir, usando $S_{ff}^{(2)}$ en vez de $\Omega^{(2)}$), aunque ésta existe. Las longitudes muestrales que elegiremos serán, esencialmente, las utilizadas por West (1996) en el ejercicio de la sección 5.2 de aquel trabajo.[31] Se realizarán 5000 repeticiones en cada caso.

Llevamos a cabo dos ejercicios de simulación diferentes. El primero de ellos lo denotaremos por Experimento B MF y responde exactamente al diseño del Experimento B del apartado anterior. Por la teoría desarrollada en West (1996) para SE, en McCracken (2000) para AE, y en el Corolario 1 de este trabajo para f discreta, ya sabemos que en una situación como ésta, con independencia entre errores de previsión simétricos y regresores, se tendrá que $F = 0$, y, por tanto, el tamaño asintótico debe coincidir con el teórico. Es decir, se tiene la IAIP, pero no conocemos si esta propiedad es aproximadamente válida también en muestras finitas, y es lo que se pretende comprobar. El segundo de los experimentos (Experimento A MF) pretende ilustrar sobre el sesgo en tamaño derivado de utilizar erróneamente $S_{ff}^{(2)}$ en vez de $\Omega^{(2)}$ cuando los errores de previsión no son independientes de los regresores y, por tanto, se sabe que $F \neq 0$. Por ello, el diseño del ejercicio es exactamente el mismo que en Experimento A, pero para muestras finitas. Este

[31] La única diferencia respecto a los tamaños muestrales elegidos en West (1996) es que nosotros incluimos el caso $P = 10$ y, en uno de los experimentos, sustituimos $P = 175$ por $P = 250$.

experimento coincide con el de la sección 5.2 de West (1996), solo que en dicho artículo se empleaba solo la función de pérdidas SE.

Los resultados obtenidos se muestran en las Tablas 3 y 4.

TABLA 3. Tamaño Empírico DM. $\alpha = 5\,\%$. Experimento B MF

		f discreta		$f = \text{SE}$	$f = \text{AE}$
R	P	DM	Mult2 (aprx)	DM	DM
25	10	10,1	4,5	8,4	8,9
	25	7,6	5,7	6,6	6,7
	50	6,0	5,1	5,8	6,5
	100	6,0	5,5	5,3	5,3
	150	5,1	4,8	5,4	4,8
	250	5,5	5,3	5,2	5,0
50	10	10,2	4,9	9,2	9,7
	25	6,3	4,5	6,9	6,5
	50	5,8	4,7	5,9	6,0
	100	5,7	5,1	5,3	5,9
	150	5,2	4,9	5,2	5,6
	250	4,7	4,5	4,8	5,2
100	10	9,8	4,5	8,8	9,5
	25	6,7	5,0	6,4	6,5
	50	5,6	4,9	5,4	5,8
	100	5,2	4,7	5,2	5,5
	150	4,9	4,7	5,2	5,2
	250	5,9	5,5	5,0	5,1

SE (AE): cuadrado (valor absoluto) del error de previsión.

TABLA 4. Tamaño Empírico DM. $\alpha = 5\,\%$. Experimento A MF

			f discreta		$f = \text{SE}$	$f = \text{AE}$
π	R	P	DM	Mult2 (aprx)	DM	DM
0,4	25	10	23,0	14,3	26,8	29,7
1		25	22,7	20,0	38,0	36,0
2		50	22,2	20,9	44,4	40,3
4		100	21,4	20,7	49,1	45,0
6		150	20,6	20,2	50,2	46,2
7		175	20,3	19,8	50,8	47,5
0,2	50	10	16,7	9,2	20,5	22,1
0,5		25	15,3	12,6	29,1	28,6
1		50	15,3	13,8	36,3	34,5
2		100	16,2	15,4	43,2	39,7
3		150	15,4	14,9	46,9	42,2
0,1	100	10	12,4	5,7	13,8	14,8
0,25		25	9,9	7,8	20,5	18,9
0,5		50	10,0	9,0	27,3	24,1
1		100	10,3	9,8	35,9	31,4

SE (AE): cuadrado (valor absoluto) del error de previsión.

Las conclusiones que se deducen de estos resultados son bastante claras:

a) En primer lugar, la propiedad de IAIP que estaba garantizada para tests como DM bajo independencia entre regresores y errores de previsión (en el contexto de modelos lineales no anidados, con variables estacionarias y simetría en la distribución de los errores) *se mantiene en muestras finitas* (véase Tabla 3), prácticamente para cualquier combinación R, P, para las tres funciones evaluadas. Solo en el caso $P = 10$ se detecta cierto sesgo en DM, pero es razonablemente pequeño. Para Mult2-aprx, el sesgo es prácticamente nulo incluso en dicho caso.

b) En segundo lugar, cuando aparece correlación entre regresores y errores de previsión, el sesgo en el tamaño de los tests es ya significativo para cualquier función de pérdidas que se emplee, pero es sistemáticamente bastante *más moderado si la función elegida es de tipo discreto*. Así, el sesgo máximo resultante en el Experimento A MF (para nivel de significación de 5 %) es de 17 % con f discreta, mientras alcanza hasta el 45 % si el test usa SE como función de pérdidas, y el 42 % si es el AE la pérdida seleccionada. Aunque 17 % es también una cuantía muy elevada como para pensar que el test puede aplicarse con f discreta sin preocuparse de corregir su varianza de la IP, probablemente proceder de dicho modo es razonable en la práctica, por el argumento que mostramos a continuación. No es habitual que el usuario pretenda hacer ejercicios de predicción estimando su modelo con menos de 100 observaciones. Si admitimos entonces que el caso verdaderamente relevante en la práctica es $R = 100$, el sesgo en tamaño (para $\alpha = 5\,\%$) es siempre inferior al 5 % usando f discreta, mientras llega a superar el 30 % y 25 % con SE y AE, respectivamente, para $P = 100$. El caso $R = P = 100$ representa la diferencia más manifiesta entre utilizar la función discreta y la continua: con la primera, la aplicación del test sin corrección alguna es perfectamente válida,[32] mientras con la segunda, estamos obligados a incorporar la compleja corrección de la varianza (construyendo $\Omega^{(2)}$), o, de lo contrario, la conclusión que se obtenga de la aplicación del test carece por completo de fiabilidad.

c) Por último, cabe mencionar que, para funciones de pérdida discretas, la versión de DM que denotamos como test Mult2-aprx parece ser más exacta en tamaño que la implementación habitual del test cuando el número de previsiones evaluado es muy pequeño.

[32] Recordemos que $R = P = 100$ significa realizar una primera estimación del modelo con 100 observaciones y generar la previsión para $t = 101$ a partir del modelo estimado y, a partir de entonces, añadir a la muestra un nuevo dato cada periodo, reestimar el modelo (con $R + t$ datos), para prever el periodo siguiente ($t = R + t + 1$), hasta un total de P periodos (y P previsiones).

4. Modelos lineales anidados (variables estacionarias)

En la sección anterior se ha abordado el problema de la IP en tests de evaluación o comparación de capacidad predictiva dentro del contexto de modelos lineales con variables estacionarias. Pero, en el caso de comparación entre previsiones de dos o más modelos, se exigía que éstos fueran no anidados. En esta sección, vamos a tratar el mismo problema pero cuando dos de los modelos de previsión son anidados. Consideraremos que dos modelos son anidados si y solo si, bajo restricciones paramétricas en *solo uno* de ellos, el otro se convierte en un caso particular del primero. Como veremos a continuación, el teorema (1a) de West (1996) no es válido en dicha situación. De hecho, la distribución asintótica del principal contraste de comparación predictiva entre dos conjuntos de previsiones, el test DM, ya no va a verificar normalidad de forma general, incluso aunque se estuviera en una situación en la que los parámetros de los modelos de previsión fueran conocidos. Es decir, en general, DM no es aplicable en su versión habitual ni en el marco de Diebold y Mariano (1995) (sin IP) ni en el de West (con IP). Sin embargo, McCracken (2007) logra derivar la distribución asintótica del estadístico DM en modelos anidados bajo un conjunto de supuestos sobre la función de pérdidas y sobre el contexto predictivo: (i) función de pérdida diferenciable, (ii) la función de pérdida para evaluar las previsiones debe ser la misma que la usada en la estimación (por ejemplo, si se usó OLS, la pérdida debe ser SE), e (iii) no existe autocorrelación en las pérdidas. La distribución asintótica obtenida por McCracken (2007) no es estándar y se formula en términos de movimientos brownianos. Nuestro interés se centra en lo que ocurre en el test DM en modelos anidados cuando la función de pérdida es de tipo discreto, incumpliendo, por consiguiente, tanto (i) como (ii). Si bien la derivación de la distribución asintótica de DM bajo pérdidas discretas queda fuera del alcance de este trabajo, comprobaremos en distintos ejercicios de simulación el comportamiento del test empleando $N(0,1)$ como distribución de contraste, tal y como hicimos en la sección anterior. Como se verá más adelante, los resultados muestran que, en caso de emplear la función discreta, el deterioro de las propiedades del test DM en el contexto de modelos anidados cuando se usa la Normal como distribución de contraste, es sensiblemente menos drástico que el que se produce en caso de que se proceda del mismo modo pero con pérdida SE, e incluso el coste podría considerarse aceptable. En cambio, en caso de utilizar SE, los pésimos resultados que tiene el uso de la distribución $N(0,1)$ obligan a aplicar los percentiles presentados por McCracken (2007).

4.1. Introducción al problema. Revisión de literatura

La aplicación del teorema de West (1a) para la comparación de predicciones procedentes de dos modelos distintos es inmediata, y el test resultante es (8), tal y como dedujimos en el apartado 2.2. Así, el test DM puede escribirse en el contexto de IP como:

$$P^{1/2}(\overline{d} - E(d_t)) \overset{a}{\sim} N(0, \Omega_d), \tag{20}$$

siendo $d_t = f_{1t} - f_{2t}$ y, por tanto, $\overline{d} = \overline{f}_1 - \overline{f}_2$, en la notación de Diebold y Mariano (1995). A diferencia de lo desarrollado en dicho trabajo, ahora $\Omega_d = E(\overline{d} - E(d_t))^2$ incluye la IP asociada a las estimaciones de los modelos, es decir, $\Omega_d = \Omega_{11} + \Omega_{22} - 2\Omega_{12}$, siendo Ω_{ij} el elemento i, j de la matriz Ω de (1b) obtenida por West (1996).

El test DM se aplicaría, por tanto, en base a la expresión:

$$P^{1/2}\overline{d}\widehat{\Omega}_d^{-1/2} \overset{a}{\sim} N(0,1), \tag{21}$$

siendo $\widehat{\Omega}_d$ una estimación consistente de Ω_d, y habiéndose utilizado la condición de la hipótesis nula $E(d_t) = 0$.

La distribución asintótica (20) se deriva de modo análogo a como se hizo para $P^{1/2}(\overline{f} - E(f_t))$ en el apartado 2.1.1.[33] Es decir, a partir de una expansión de Taylor sobre $P^{1/2}\overline{d} = P^{-1/2}\sum_{t=R}^{T}(f_{1t}(\widehat{\beta}_{1t}) - f_{2t}(\widehat{\beta}_{2t}))$:

[33] En realidad, se trata de multiplicar por el vector $(1, -1)$ la aproximación de West (1996) del estadístico vectorial $P^{1/2}(\overline{f} - E(f_t))$.

$$P^{1/2}\overline{d} = P^{-1/2}\sum_{t=R}^{T}(f_{1t}(\beta_1^*) - f_{2t}(\beta_2^*)) + P^{-1/2}\sum_{t=R}^{T}(\left[\frac{\partial f_1}{\partial \beta_1}\right]_{\beta_1^*}(\widehat{\beta}_{1t} - \beta_1^*) - \left[\frac{\partial f_2}{\partial \beta_2}\right]_{\beta_2^*}(\widehat{\beta}_{2t} - \beta_2^*)) + o_p(1). \quad (22)$$

En modelos no anidados, el primer término genera la varianza asintótica correspondiente a la condición de variable aleatoria de los datos y regresores (incertidumbre no paramétrica), y el segundo, la asociada a la estimación de los modelos (IP). Sin embargo, cuando los modelos están anidados ocurre lo siguiente:

a) Bajo H_0, el primer término de (22) se anula para todo t, porque $f_{1t}(\beta_1^*) = f_{2t}(\beta_2^*)$.

b) Bajo H_0, el segundo término de (22) converge en probabilidad a cero.

Como consecuencia de a) y b), se tiene que $P^{1/2}\overline{d} \xrightarrow{p} 0$.

c) Pero, además, usando resultados en West (1996), puede mostrarse que $\Omega_d = 0$. Esencialmente, lo que ocurre es que $f_{1t}(\widehat{\beta}_{1t})$ y $f_{2t}(\widehat{\beta}_{2t})$ son asintóticamente iguales, y, por tanto, la varianza asintótica asociada a $\overline{f}_1$ y a $\overline{f}_2$ es la misma ($\Omega_{11} = \Omega_{22}$), mientras que f_{1t} y f_{2t} están perfectamente correlados asintóticamente ($\Omega_{12} = \Omega_{11}^{1/2}\Omega_{22}^{1/2}$). De este modo, se cumple que $\Omega_d = \Omega_{11} + \Omega_{22} - 2\Omega_{12} = 0$. Finalmente, por consistencia de $\widehat{\Omega}_d$, se tiene que $\widehat{\Omega}_d \xrightarrow{p} 0$.

Por lo tanto, tanto numerador como denominador del estadístico en (21) convergen en probabilidad a cero, por lo que la distribución de éste podría ser divergente, degenerada o convergente. Desde luego, no será asintóticamente normal, a diferencia de lo afirmado en (21) para modelos no anidados. Véase que el problema surgiría igualmente si no existiera IP.

McCracken (2007) aporta una solución al problema. Aunque el argumento habitual de aproximación de Taylor no permite hallar la distribución asintótica de $P^{1/2}\overline{d} = P^{-1/2}\sum_{t=R}^{T}(f_{1t}(\widehat{\beta}_{1t}) - f_{2t}(\widehat{\beta}_{2t}))$, McCracken (2007) logra derivarla siempre que se verifiquen los supuestos (i)-(iii) mencionados en la introducción previa a este apartado. Para ello, obtiene la distribución de los estadísticos D_1 y D_2, y, finalmente, del ratio $\frac{D_1}{D_2}$, siendo:

$D_1 = \sum_{t=R}^{T}(f_{1t}(\widehat{\beta}_{1t}) - f_{2t}(\widehat{\beta}_{2t}))$, $D_2 = \sum_{t=R}^{T}(f_{1t}(\widehat{\beta}_{1t}) - f_{2t}(\widehat{\beta}_{2t}))^2$ y $\frac{D_1}{D_2} = P^{1/2}\overline{d}\widehat{\Omega}_d^{-1/2}$.

Es decir, obtiene la distribución asintótica del estadístico DM en el contexto de modelos anidados bajo el cumplimiento de los supuestos (i)-(iii). La distribución obtenida es función de movimientos brownianos. McCracken (2007) obtiene por simulación los percentiles de la distribución encontrada, que dependen del valor del parámetro $\pi = \lim_{T\to\infty} P/R$ y de un parámetro k, que representa el número de parámetros en los que el modelo "largo" excede al "corto". Entiéndase que el estadístico de contraste no incorpora IP, sino que es la distribución la que varía según el grado de ésta, medida a través del parámetro π. La distribución asintótica obtenida difiere notablemente de $N(0,1)$, tanto en su apuntamiento como en la posición que ocupa en la recta real. En concreto, se encuentra desplazada hacia la izquierda respecto a la de $N(0,1)$ (ie, es asimétrica respecto de cero y la probabilidad de que el estadístico DM tome valor negativo es mucho mayor de 0,5). Su asimetría respecto de cero crece con π y k, y su apuntamiento, con π. Únicamente en el caso $\pi = 0$, la distribución de $\frac{D_1}{D_2}$ es $N(0,1)$. Esto significa que el test DM conserva normalidad asintótica en un contexto de IAIP con modelos anidados, pero no de forma general, sino solo bajo el conjunto de restricciones (i)-(iii).

Las limitaciones del resultado logrado en McCracken (2007) se encuentran, por un lado, en el hecho de que la distribución derivada no es estándar, lo que genera cierta incomodidad en la implementación práctica del test, y, por otro, en la exigencia de tres supuestos relativamente restrictivos. Debido a ellos, en la práctica, el test de McCracken (2007) se podrá aplicar esencialmente si concurren simultáneamente las siguientes condiciones: la función de pérdida es SE, los parámetros se estimaron por OLS,[34] y el horizonte de previsión es uno (en cuyo caso el supuesto (iii) puede considerarse satisfecho, y, de este modo, el estadístico $P^{-1}D_2$ será efectivamente un estimador consistente de Ω_d).

Por supuesto, además de la versión del test DM presentada por McCracken (2007), que McCracken denomina OOS-t test, existen otros métodos alternativos para contrastar igualdad de capacidad predictiva entre dos modelos de previsión anidados. De hecho, en McCracken (2007) se propone también un estadístico

[34] Siempre que hablamos del método OLS, en realidad, se están incluyendo todos los estimadores que minimizan la misma función que OLS, por ejemplo, NLLS o Máxima Verosimilitud si se presupone normalidad en las perturbaciones.

tipo F para el mismo propósito y se deriva su distribución no estándar (OOS-F test). Por su parte, Clark y McCracken (2001) obtienen las distribuciones asintóticas, todas ellas no estándar, para tres tests de "encompassing" adaptados para incorporar la IP en el contexto de modelos anidados y bajo el supuesto de horizonte de previsión uno: ENC-T y ENC-REG, cuyas versiones originales se deben a Harvey et al (1998) y Ericsson (1992), respectivamente, y un tercer test, ENC-NEW, propuesto por los autores en el artículo. Se evalúan las propiedades en muestras finitas de estos tres contrastes y del test OOS-t de McCracken (2007), a través de ejercicios de simulación, y se estiman también sus tamaños y potencias cuando se emplea $N(0,1)$ como distribución de contraste, para comprobar el sesgo que se genera en tal caso. Finalmente, Clark y West (2007) proponen un test similar a OOS-t pero solo para función de pérdida cuadrática y sustituyendo el cálculo del habitual MSE por un estadístico "MSE-ajustado", y la distribución asintótica obtenida resulta presentar un grado de asimetría hacia la izquierda (respecto de cero) sensiblemente menor que el del test OOS-t. Aunque la distribución teórica resulta no estándar, Clark y West (2007) realizan experimentos de simulación implementando el test con $N(0,1)$ como distribución de contraste. Los sesgos generados en tamaño bajo esta forma de proceder son razonablemente pequeños, en muestras finitas. En concreto, se obtienen tamaños empíricos en el intervalo (5,5 %, 9 %) en todos los casos examinados, cuando el tamaño teórico era 10 %.

Obviamente, no existen resultados para el test DM en el contexto de modelos anidados cuando la función de pérdida es de tipo discreto (incumpliendo, por tanto, las condiciones exigidas por el teorema de McCracken (2007) para el test OOS-t). Ni siquiera existe teoría estadística al respecto para funciones de pérdida no diferenciables mucho más habituales, como, por ejemplo, AE. Derivar una distribución asintótica en el caso de la función de pérdida discreta queda fuera del alcance de este trabajo, pero consideramos muy relevante chequear las propiedades del test DM en esta situación, implementando el contraste con (21), es decir, usando la habitual distribución Normal, aun a sabiendas de que no es ésta la distribución asintótica correcta. Con esta intención llevaremos a cabo ejercicios de simulación, implementando el test DM tanto con f discreta como con SE, pero empleando la distribución $N(0,1)$ para aplicar los contrastes.[35]

Llevaremos a cabo tres experimentos con modelos anidados. Además, vamos a considerar otra categoría de modelos de previsión similares a los anidados, pero con algunas diferencias. Se trata de modelos que son iguales bajo la hipótesis nula, pero entre los que no existe anidamiento. En este caso, se necesitan restricciones paramétricas en los *dos* modelos para que se conviertan en el mismo, a diferencia de lo que ocurre en los anidados. Nos referimos, por ejemplo, a modelos de tipo $y_t = Z_t'\beta^* + x_{1t}\gamma_1^* + u_{1t}$ e $y_t = Z_t'\beta^* + x_{2t}\gamma_2^* + u_{2t}$, siendo $V(x_{1t}) = V(x_{2t})$ y, bajo H_0, $\gamma_1^* = \gamma_2^* = 0$. En este tipo de casos, es obvio que se verifica también la condición $f_{1t}(\beta_1^*) = f_{2t}(\beta_2^*)\ \forall t$, y que aplica el mismo razonamiento que se hizo arriba para explicar el problema del test DM en modelos anidados. En cualquier caso, la relevancia de los resultados acerca de esta nueva categoría de modelos será inferior a la de los anidados, puesto que la frecuencia con la que aparece esta situación en la práctica es mucho menor que la de los otros.[36]

4.2. Ejercicios de simulación

Como se acaba de mencionar, realizaremos ejercicios de simulación para modelos anidados, y, además, para esa otra clase de modelos similares a los anidados, que se han comentado al final del apartado anterior. Llamaremos a esta segunda categoría "modelos no exactamente anidados". Los experimentos solamente estarán enfocados a estimar el tamaño de los tests involucrados, de modo que la hipótesis nula será cierta en todos ellos. Cada uno de los experimentos se realizará en dos versiones: para estimar el tamaño de los contrastes en muestras finitas (para distintos valores de π), y los análogos pero para estimar el tamaño asintótico. Para este último propósito, se emplearán valores de T suficientemente grandes como para considerar válido el resultado, pero suficientemente pequeños como para que el coste computacional sea razonable. Además de los tamaños estimados, se ofrecerán las funciones de densidad empíricas obtenidas en los ejercicios para los estadísticos de contraste.

[35] También aplicaremos el test Mult2-aprx presentado en el capítulo 2, para comprobar si produce alguna mejoría respecto a los resultados de DM con f discreta.

[36] En el proceso de construcción de modelos predictivos, es habitual que la decisión para añadir un regresor descanse en un test de comparación de capacidad predictiva que enfrente el modelo inicial con otro igual, solo que incluyendo el regresor adicional. Por tanto, bajo la hipótesis nula de que el último regresor no tiene capacidad explicativa, se trata de dos modelos anidados. Esto justifica la relevancia de disponer de un test con buenas propiedades para dicho contexto.

4.2.1. Diseño de los experimentos

PGD y Modelos de previsión

Modelos anidados Llevaremos a cabo tres experimentos:

1. Modelos lineales de regresión (Experimento 1): El PGD es $y_t = z_{1t} + \delta^* z_{2t} + u_t$, siendo $u_t \overset{iid}{\sim} N(0,1)$, mientras los modelos que compiten para prever son $y_t = \beta_{11}^* z_{1t} + u_{1t}$ (M_1, modelo restringido) e $y_t = \beta_{21}^* z_{1t} + \beta_{22}^* z_{2t} + u_{2t}$ (M_2, modelo sin restringir). La hipótesis nula es $H_0 \equiv E(f_{1t}(\beta_1^*)) \leq E(f_{2t}(\beta_2^*))$, y la alternativa, $H_1 \equiv E(f_{1t}(\beta_1^*)) > E(f_{2t}(\beta_2^*))$ ($\beta_1^* = \beta_{11}^*$ y $\beta_2^* = (\beta_{21}^*, \beta_{22}^*)'$). En términos paramétricos, estas hipótesis equivalen a $H_0 \equiv \delta^* = 0$ y $H_1 \equiv \delta^* \neq 0$. Fijaremos $\delta^* = 0$, de modo que H_0 sea cierta y M_1 y M_2 sean modelos anidados.

2. Modelos VAR (Experimento 2): Éste es el mismo ejercicio realizado en McCracken (2007) y en Clark y McCracken (2001). El PGD es de la forma:

$$\begin{pmatrix} y_t \\ x_t \end{pmatrix} = \begin{pmatrix} 0{,}3 & \delta^* \\ 0 & 0{,}5 \end{pmatrix} \begin{pmatrix} y_{t-1} \\ x_{t-1} \end{pmatrix} + \begin{pmatrix} u_{y,t} \\ u_{x,t} \end{pmatrix}, \tag{23}$$

siendo $(u_{y,t}, u_{x,t})' \overset{iid}{\sim} N(0_{2\times 1}, I_{2\times 2})$.

Se pretende prever solamente la variable y_t. Para ello, compiten dos modelos de previsión: un AR(1) sobre y_t (M_1, modelo restringido) y un VAR(1) sobre el vector $(y_t, x_t)'$ (M_2, modelo sin restringir). Las hipótesis nula y alternativa son exactamente las mismas que en el Experimento 1. Nuevamente, fijaremos $\delta^* = 0$, de modo que la hipótesis nula sea verdadera y los modelos de previsión resulten anidados.

3. Modelos con raíz unitaria (Experimento 3): El PGD es $y_t = \delta^* y_{t-1} - 0{,}5 y_{t-2} + u_t$, siendo $u_t \overset{iid}{\sim} N(0,1)$, y verificándose $\delta^* > -1{,}5$.[37] Se pretende prever ∇y_t. Para ello, compiten dos modelos de previsión: un ARIMA(1,1,0), es decir, un AR(1) especificado sobre ∇y_t (M_1, modelo restringido) frente a un ARIMA(2,0,0), es decir, un AR(2) sobre y_t (M_2, modelo sin restringir), cuyas previsiones en niveles se usarán para construir las previsiones en diferencias. En términos de capacidad predictiva de los modelos, las hipótesis vuelven a ser las especificadas en el Experimento 1, mientras sus equivalencias paramétricas son ahora $H_0 \equiv \delta^* = 1{,}5$ y $H_1 \equiv -1{,}5 < \delta^* < 1{,}5$. De este modo, bajo la hipótesis nula, el PGD presenta una raíz unitaria, y M_1 y M_2 serían modelos anidados. Dado que se trata de un ejercicio para estimar el tamaño de los tests, fijaremos $\delta^* = 1{,}5$.

Modelos "no exactamente anidados" Presentamos un solo experimento de este tipo:

4. Modelos lineales de regresión con variables sin valor explicativo (Experimento 4): En este caso, el PGD es $y_t = z_t + \delta_1^* x_{1t} + \delta_2^* x_{2t} + u_t$, siendo $u_t, x_{1t}, x_{2t} \overset{iid}{\sim} N(0,1)$. Por su parte, los modelos para prever son $y_t = \beta_{10}^* z_t + \beta_{11}^* x_{1t} + u_{1t}$ (M_1) e $y_t = \beta_{20}^* z_t + \beta_{22}^* x_{2t} + u_{2t}$ (M_2). En términos de habilidad predictiva de los modelos, podemos escribir las hipótesis como $H_0 \equiv E(f_{1t}(\beta_1^*)) = E(f_{2t}(\beta_2^*))$ frente a $H_1 \equiv E(f_{1t}(\beta_1^*)) \neq E(f_{2t}(\beta_2^*))$. Nótese que ahora la hipótesis alternativa no establece un orden de preferencia entre los modelos, a diferencia de lo que ocurría en los tres experimentos anteriores. Paramétricamente, las hipótesis se pueden escribir como $H_0 \equiv \delta_1^* = \delta_2^* = 0$ y $H_1 \equiv \neg H_0$. Es decir, bajo H_0, los dos modelos incluyen una variable (cada uno una distinta) sin valor explicativo, de modo que sus previsiones no serán iguales, pero, en cambio, se verificará $f_{1t}(\beta_1^*) = f_{2t}(\beta_2^*)$ asintóticamente, como en modelos anidados ($\beta_1^* = (\beta_{10}^*, \beta_{11}^*)'$ y $\beta_2^* = (\beta_{20}^*, \beta_{22}^*)'$)). Fijaremos $\delta_1^* = \delta_2^* = 0$, de modo que H_0 sea cierta, y, por tanto, los modelos M_1 y M_2 tengan la misma capacidad predictiva.

Resto del diseño El resto de especificaciones del diseño es común a los cuatro experimentos:

a) Estimación y previsiones: los parámetros se estiman por OLS, y las previsiones se realizan exclusivamente a horizonte uno. El esquema de estimación empleado es el que se conoce como "recursivo", es decir, la estimación en t se realiza con los datos correspondientes al intervalo $[1, t]$, incorporando, por tanto, una observación nueva cada periodo, desde $t = R$ hasta $t = T = R + P - 1$, siendo $T + 1$ el tamaño de la muestra total y P el número de previsiones realizadas. No se utiliza ninguno de los otros dos esquemas

[37] De este modo, se cumplen con seguridad dos de las tres condiciones de estacionariedad de un proceso estocástico AR(2), de parámetros ϕ_1 y ϕ_2: $|\phi_2| < 1$ y $\phi_2 - \phi_1 < 1$. La tercera condición de estacionariedad, pendiente de verificar, es: $\phi_2 + \phi_1 < 1$.

posibles, "rolling" y "fixed". Al realizar la previsión en t para el periodo $t+1$, se suponen conocidos todos los regresores del modelo en $t+1$.

b) Tests: una vez se tienen los conjuntos de previsiones y los verdaderos datos asociados a los periodos de previsión, se ejecutarán tres contrastes: por un lado, el test DM, en dos versiones, bajo función de pérdida f discreta y bajo el SE, y, por otro, el test Mult2-aprx, bajo la misma f discreta que en DM. Como en experimentos de la sección 3, la función de pérdida discreta que usamos es una definida (a) por la partición $P_N = \{l_0 = -\infty, l_1 = -0{,}25\sigma_w, l_2 = 0, l_3 = +0{,}25\sigma_w, l_4 = +\infty\}$, siendo σ_w la desviación típica de la variable w_t sobre que interesa prever ($w_t = y_t$ en los Experimentos 1, 2 y 4, mientras $w_t = \nabla y_t$ en el Experimento 3), y (b) por la matriz (9).

Los tres contrastes se implementarán utilizando $N(0,1)$ como distribución de contraste, aun a sabiendas de que ésta no es la distribución correcta en el caso de modelos anidados para el test DM bajo SE, tal y como se demuestra en McCracken (2007), y con la sospecha de que tampoco lo será cuando f es de tipo discreto. Tal y como se explicó en 4.1, nuestro interés está precisamente en comprobar cuánto se resienten las propiedades del test DM (o Mult2-aprx) en un contexto de modelos anidados y f discreta cuando se implementa como si el asunto del anidamiento no tuviera efecto en su distribución asintótica. Dado que la verdadera distribución no se conoce, resultaría tranquilizador encontrar que los sesgos en tamaño del test DM (o Mult2-aprx) cuando se implementa usando la $N(0,1)$ no son grandes.

Deben hacerse dos comentarios añadidos sobre la aplicación de los contrastes en estos ejercicios. Por un lado, al utilizar una función discreta, y dado que M_1 y M_2 hacen previsiones similares, cuando P es pequeño, se producen ocasionalmente realizaciones del experimento en las que los P diferenciales de pérdidas ($d_t = f_{2t}(\widehat{\beta}_2) - f_{1t}(\widehat{\beta}_1)$) resultan nulos. En tales casos, aunque el numerador y el denominador del estadístico de contraste de DM (véase (21)) será cero y el contraste, teóricamente, no puede implementarse, procederemos directamente a no rechazar H_0, por razones obvias. Por otro lado, en el Capítulo 2 se demostró que, bajo la hipótesis nula de igual capacidad predictiva de los dos conjuntos de previsión, los tests Mult2-aprx y DM (con f discreta) son asintóticamente equivalentes, siempre que las pérdidas no estén autocorreladas, tal y como ocurre en estos experimentos.[38] Por ello, veremos cómo los resultados de tamaño de ambos tests son prácticamente iguales, salvo en casos en que el número de previsiones P sea muy pequeño.

Finalmente, dadas las especificaciones de la hipótesis nula en cada ejercicio, es obvio que los tests deben implementarse con una sola región crítica de probabilidad igual al nivel de significación α en el caso de los Experimentos 1, 2 y 3, y con dos regiones críticas, cada una de probabilidad $\alpha/2$, en el caso del Experimento 4.

c) Longitudes muestrales P, R: en los ejercicios de muestras finitas, se usarán las mismas combinaciones P, R de los experimentos de Monte Carlo realizados en Clark y McCracken (2001), a saber: se emplea $R = 50$ con $P = 100$, 150 y 250; $R = 100$ con $P = 10$, 20, 40, 100 y 200; y, finalmente, $R = 200$ con $P = 20$, 40, 80 y 200. Estas combinaciones se corresponden con una gama de valores de $\widehat{\pi} = P/R$ habituales en la práctica: 0,1, 0,2, 0,4, 1, 2, 3 y 5.

Como ya se dijo arriba, nos interesa estimar también el tamaño asintótico de los tests. Para tal propósito, realizamos los mismos ejercicios pero con valores de P y R mucho mayores que los anteriores, pero conservando los mismos ratios $\widehat{\pi}$ que en muestras finitas: $R = 10000$ y $P = 1000$ para $\widehat{\pi} = 0{,}1$; $R = 5000$ y $P = 1000$ para $\widehat{\pi} = 0{,}2$; $R = 2500$ y $P = 1000$ para $\widehat{\pi} = 0{,}4$; $R = 1000$ y $P = 1000$ para $\widehat{\pi} = 1$; $R = 1000$ y $P = 2000$ para $\widehat{\pi} = 2$; $R = 500$ y $P = 1500$ para $\widehat{\pi} = 3$; y, finalmente, $R = 500$ y $P = 2500$ para $\widehat{\pi} = 5$.

d) Finalmente, el nivel de significación será siempre $\alpha = 10\,\%$, y el número de repeticiones de cada experimento será 5000 en los ejercicios de muestras finitas, y 10000 en los destinados a estimar tamaño asintótico.

[38] Ésta es la Propiedad b del caso 1 del apartado "Conclusiones sobre la comparación entre Mult2-aprx y DM en tamaño y potencia" de la sección donde se presenta el test Mult2-aprx, en el capítulo 2.

4.2.2. Resultados

En las Tablas 5 a 8 se presentan los tamaños estimados en muestras finitas para los tests en los cuatro experimentos descritos arriba,[39] mientras las Tablas 9 a 12 ofrecen estimaciones sobre tamaño asintótico. Además, en las Figuras 1 a 7 se adjuntan las funciones de densidad empíricas obtenidas en los ejercicios asintóticos,[40] para poder ilustrar sus diferencias respecto a la distribución $N(0,1)$. Las conclusiones que se deducen son las siguientes:

Modelos anidados (Experimentos 1 a 3)

a) Las dos implementaciones de DM (tanto con f cuadrática como con f discreta) y Mult2-aprx infraestiman tamaño al aplicarse con $N(0,1)$ como distribución de contraste. En coherencia con la teoría presentada por McCracken (2007) para el caso de funciones de pérdida diferenciables,[41] el sesgo aumenta con π.

b) En términos asintóticos, las diferencias entre usar f discreta y cuadrática (SE) en el test DM con distribución de contraste $N(0,1)$ son muy notables. Promediando a lo largo de todos los valores de π y de los tres experimentos, el tamaño de DM (o Mult2-aprx) bajo f discreta es de 8,6 %, por 3,1 %, en el caso del SE. Salvo en los casos de $\pi \leq 1$ en el Experimento 3, donde las discrepancias son menores, las diferencias entre el tamaño con f discreta y cuadrática en el resto de diseños siempre están entre cinco y ocho puntos, y aumentan con π. Mientras la infraestimación de tamaño usando SE (y $N(0,1)$, recordemos) es muy grave, el empleo de función de pérdida discreta garantiza tamaños asintóticos próximos al teórico, siendo el sesgo superior a dos puntos solo en niveles de $\pi \geq 2$ (y únicamente en los Experimentos 2 y 3), situaciones en las que, a su vez, el test DM bajo f cuadrática prácticamente nunca rechaza H_0.

c) En muestras finitas, como era de esperar, los tamaños de DM bajo f discreta empeoran, pero siguen siendo razonables, con niveles entre 5 % y 9 % en prácticamente todos los ejercicios realizados. En cualquier caso, estos resultados son de nuevo mucho mejores que los que se obtendrían empleando SE como función de pérdida, con la que incluso se producen tamaños inferiores a 1 % en casi todas las situaciones de $\pi = 3$ y $\pi = 5$. Promediando los tamaños obtenidos en todas las combinaciones (P, R) de los tres experimentos, el tamaño medio de DM (y Mult2-aprx) en muestras finitas es 7,1 %, por 4,2 % con SE.

d) Obviamente, los sesgos de tamaño obtenidos se producen por la diferencia entre la verdadera distribución asintótica de los estadísticos y la distribución $N(0,1)$ empleada. En las Figuras 1 a 7 pueden apreciarse tales diferencias. En dichos gráficos se presentan histogramas con las densidades empíricas de las 10000 realizaciones del estadístico DM generadas en cada experimento asintótico para la función de pérdida discreta[42] y para el SE, respectivamente, junto con la densidad de una $N(0,1)$. Por coherencia con el Capítulo 2 de esta tesis, hemos definido el estadístico DM (así como el correspondiente a Mult2-aprx) en base al diferencial $\overline{f}_2 - \overline{f}_1$, en vez de $\overline{f}_1 - \overline{f}_2$, de modo que la cola de rechazo de los tests es la izquierda. En el caso del estadístico DM bajo SE, la función de densidad asintótica correcta se encuentra siempre desplazada a la derecha respecto a la de la $N(0,1)$,[43] de modo que la probabilidad de que el estadístico sea inferior al percentil α de una $N(0,1)$ es muy inferior a α, resultado que ya fue apuntado por McCracken (2007). En cambio, en el caso de usar f discreta, no se dispone de teoría sobre la densidad asintótica verdadera del estadístico DM. Los tres primeros gráficos de las Figuras 1 a 7 nos aportan evidencia a este respecto. De acuerdo a ellos, parece que dicha distribución es mucho menos asimétrica que la correspondiente a la implementación de DM con SE, y su apuntamiento solo difiere del de $N(0,1)$ en niveles de $\pi \leq 0{,}4$. En general, la densidad discrepa de la $N(0,1)$ bastante menos que bajo pérdida SE.

[39] Tanto McCracken (2007) como Clark y McCracken (2001) realizan exactamente el ejercicio de Monte Carlo que nosotros hemos denominado Experimento 2 (pero solo en muestras finitas), utilizando la distribución $N(0,1)$ para el contraste. Clark y McCracken (2001) emplean el mismo tamaño teórico que nosotros, $\alpha = 10\,\%$. Tal y como era esperable, sus resultados y los que presentamos nosotros en este documento son prácticamente iguales.

[40] Tanto en las Tablas como en las Figuras, se emplea la notación MF para referirse a los ejercicios de muestras finitas y AS, para los de carácter asintótico (de muestras muy largas, en realidad).

[41] Recordemos que, según McCracken (2007), la distribución asintótica verdadera del estadístico OOS-t en modelos anidados, presentaba un desplazamiento en la recta real respecto a la $N(0,1)$ creciente con π.

[42] Las densidades empíricas obtenidas para el test Mult2-aprx son idénticas a las correspondientes a DM bajo f discreta (tal y como debía ocurrir, dada la convergencia asintótica entre ambos estadísticos), por lo que no se presentan en los gráficos.

[43] De haber mantenido la definición habitual, el desplazamiento sería a la izquierda, como en McCracken (2007).

e) Analizando con más detalle los histogramas presentados, se obtienen los siguientes patrones:

– Se confirma en todos los experimentos de modelos anidados que la distribución del estadístico DM bajo SE es claramente asimétrica respecto de cero y más apuntada que la $N(0,1)$, acrecentándose ambos síntomas con π, tal y como se afirmaba en McCracken (2007).

– En cambio, si se usa una función de pérdida discreta en la implementación del test, la distribución asintótica presenta asimetría creciente con π pero el exceso de apuntamiento en relación al de la distribución normal es decreciente respecto a dicho parámetro.

– Con f discreta, la asimetría es prácticamente nula para $\pi \leq 0{,}4$, aunque las discrepancias entre la forma de la distribución y la $N(0,1)$ son muy notables. Sin embargo, tal discrepancia afecta escasamente a las colas de la distribución, lo que explica los buenos resultados del tamaño del test obtenidos para estos valores de π (Tablas 9 a 11).

– Cuando $\pi \geq 1$, el apuntamiento de la densidad asintótica de DM con f discreta coincide aproximadamente con el de la $N(0,1)$, y las formas de ambas distribuciones prácticamente son iguales. Sin embargo, se observa cierto desplazamiento a la derecha de la densidad del estadístico DM respecto a la $N(0,1)$, pero muy inferior al que presenta la distribución del estadístico cuando se implementó con pérdida SE. Dicho desplazamiento es el causante del sesgo en el tamaño de DM con f discreta que se había identificado en las Tablas 9 a 11 para los casos $\pi \geq 1$.

Modelos no exactamente anidados (Experimento 4)

El ejercicio realizado para este tipo de modelos ofrece resultados algo mejores para los tamaños de todos los tests que en el caso de modelos anidados. En términos asintóticos, DM prácticamente no presenta ningún sesgo en tamaño si se implementa con f discreta: en promedio de todos los niveles de π elegidos, el tamaño es de 9,7 %, permaneciendo todos los valores entre 9,0 % y 10,2 %. En cambio, bajo SE, el tamaño medio de DM es solo de 5,0 %, con infraestimaciones muy graves en todos los ejercicios en los que $\pi > 1$ (3,3 %, 2,4 % y 1,6 % en los casos $\pi = 2$, $\pi = 3$ y $\pi = 5$, respectivamente).

Nuevamente, las densidades empíricas correspondientes al Experimento 4 dibujadas en las Figuras 1 a 7 permiten justificar estos resultados. La distribución del estadístico DM en este tipo de modelos no exactamente anidados es simétrica respecto de cero, tanto bajo f discreta como bajo SE. Sin embargo, respecto al grado de apuntamiento, existen diferencias notables entre ambos casos. Con f discreta, el patrón es el mismo que en modelos anidados: exceso de apuntamiento respecto a la $N(0,1)$ solo en casos de $\pi \leq 0{,}4$, pero, incluso en estas situaciones, las colas siguen siendo aproximadamente igual de gruesas que las de la $N(0,1)$, lo que explica los buenos resultados de tamaño expuestos en la Tabla 12. Por el contrario, con pérdida cuadrática, el exceso de apuntamiento es más grave aún que en el caso de modelos anidados, especialmente para casos donde $\pi \geq 1$, y las colas de la distribución son mucho más finas que las de $N(0,1)$. Esto explica que, para dichos valores de π, se infraestime tamaño (Tabla 12) casi en la misma cuantía que en los experimentos de modelos anidados, pese a que la distribución de DM bajo SE sea ahora simétrica.

Conclusión final

Según los resultados obtenidos, la distribución asintótica del test DM es notablemente más robusta al anidamiento de los modelos si el contraste se implementa con una función de pérdida discreta que si se emplea el habitual SE. De utilizar este último tipo de pérdida, es obligado que, cuando el usuario compare modelos anidados, implemente el contraste DM utilizando las tablas de percentiles proporcionadas por McCracken (2007) para la verdadera distribución asintótica del test, con las incomodidades que esto conlleve. Además, si los modelos son del tipo que hemos llamado "no exactamente anidados", ni siquiera dichos percentiles son válidos. Si, por el contrario, interesa una función de pérdida discreta, es razonable seguir implementando el test DM exactamente como en (21), es decir, empleando $N(0,1)$ como distribución de contraste, independientemente de que se comparen modelos anidados, no anidados o "no exactamente anidados", aunque existirá cierto coste en la fiabilidad del test si se procede de este modo en el caso de los primeros.

TABLA 5. Tamaño Empírico DM. $\alpha = 10\,\%$. Experimento 1 MF

			f discreta		$f =$ SE
R	P	π	DM	Mult2 (aprx)	DM
50	100	2	6,9	6,8	1,0
	150	3	6,6	6,5	0,7
	250	5	7,0	6,7	0,5
100	10	0,1	3,5	3,5	9,5
	20	0,2	6,0	6,0	6,6
	40	0,4	8,3	8,1	4,4
	100	1	8,7	8,6	2,3
	200	2	8,6	8,5	1,3
200	20	0,1	4,8	4,8	7,3
	40	0,2	7,0	6,9	5,0
	80	0,4	7,9	7,9	3,6
	200	1	8,5	8,5	1,7

SE: cuadrado del error de previsión; MF: muestras finitas.

TABLA 6. Tamaño Empírico DM. $\alpha = 10\,\%$. Experimento 2 MF

			f discreta		$f =$ SE
R	P	π	DM	Mult2 (aprx)	DM
50	100	2	6,0	5,9	1,3
	150	3	5,0	5,0	0,7
	250	5	4,9	4,8	0,5
100	10	0,1	7,2	7,1	9,2
	20	0,2	7,5	7,3	6,2
	40	0,4	8,2	7,6	4,2
	100	1	7,1	7,0	2,1
	200	2	6,4	6,2	1,1
200	20	0,1	7,7	7,5	7,8
	40	0,2	8,5	8,0	5,7
	80	0,4	9,1	9,0	3,5
	200	1	6,9	6,9	1,8

SE: cuadrado del error de previsión; MF: muestras finitas.

TABLA 7. Tamaño Empírico DM. $\alpha = 10\,\%$. Experimento 3 MF

			f discreta		$f = \text{SE}$
R	P	π	DM	Mult2 (aprx)	DM
50	100	2	5,3	5,2	2,3
	150	3	5,2	5,2	1,4
	250	5	4,3	4,2	0,5
100	10	0,1	6,7	6,6	9,8
	20	0,2	8,4	8,1	8,3
	40	0,4	8,5	8,0	6,2
	100	1	7,0	6,8	3,7
	200	2	6,0	5,9	2,3
200	20	0,1	8,2	8,0	9,8
	40	0,2	9,3	8,8	8,5
	80	0,4	9,1	9,1	6,4
	200	1	7,7	7,6	4,0

SE: cuadrado del error de previsión; MF: muestras finitas.

TABLA 8. Tamaño Empírico DM. $\alpha = 10\,\%$. Experimento 4 MF

			f discreta		$f = \text{SE}$
R	P	π	DM	Mult2 (aprx)	DM
50	100	2	10,1	9,3	3,4
	150	3	9,5	8,9	2,7
	250	5	9,3	9,2	2,0
100	10	0,1	2,2	1,7	14,0
	20	0,2	6,4	5,1	10,4
	40	0,4	9,9	7,8	7,7
	100	1	10,0	8,8	5,3
	200	2	9,9	9,8	3,1
200	20	0,1	3,8	3,1	11,0
	40	0,2	8,1	6,5	9,0
	80	0,4	10,1	8,6	7,1
	200	1	9,1	9,1	4,6

SE: cuadrado del error de previsión; MF: muestras finitas.

TABLA 9. Tamaño Empírico DM. $\alpha = 10\,\%$. Experimento 1 AS

	f discreta		$f = \text{SE}$
π	DM	Mult2 (aprx)	DM
0,1	9,8	9,8	5,1
0,2	10,4	10,4	4,1
0,4	10,0	10,0	3,1
1	9,5	9,4	1,8
2	8,2	8,2	0,9
3	8,1	8,1	0,6
5	8,2	8,2	0,4

SE: cuadrado del error de previsión; AS: caso asintótico.

TABLA 10. Tamaño Empírico DM. $\alpha = 10\,\%$. Experimento 2 AS

	f discreta		$f = \text{SE}$
π	DM	Mult2 (aprx)	DM
0,1	10,4	10,4	4,9
0,2	9,3	9,3	4,3
0,4	9,0	9,0	2,8
1	7,7	7,7	1,3
2	7,7	7,7	1,1
3	6,8	6,8	0,7
5	6,4	6,4	0,3

SE: cuadrado del error de previsión; AS: caso asintótico.

TABLA 11. Tamaño Empírico DM. $\alpha = 10\,\%$. Experimento 3 AS

	f discreta		$f = \text{SE}$
π	DM	Mult2 (aprx)	DM
0,1	10,6	10,6	9,4
0,2	10,0	9,9	8,6
0,4	9,3	9,3	7,3
1	8,5	8,5	4,1
2	7,2	7,2	2,1
3	6,4	6,4	1,5
5	6,4	6,4	0,7

SE: cuadrado del error de previsión; AS: caso asintótico.

TABLA 12. Tamaño Empírico DM. $\alpha = 10\,\%$. Experimento 4 AS

	f discreta		$f = \text{SE}$
π	DM	Mult2 (aprx)	DM
0,1	9,9	9,9	9,1
0,2	9,8	9,8	7,7
0,4	9,8	9,8	6,7
1	9,0	9,0	4,5
2	9,7	9,7	3,3
3	9,9	9,9	2,4
5	10,2	10,2	1,6

SE: cuadrado del error de previsión; AS: caso asintótico.

FIGURA 1. DENSIDAD EMPÍRICA ASINTÓTICA ESTADÍSTICO DM BAJO IP
FUNCIONES PÉRDIDA: SE Y DISCRETA. MODELOS ANIDADOS. CASO $\pi = 0,1$.

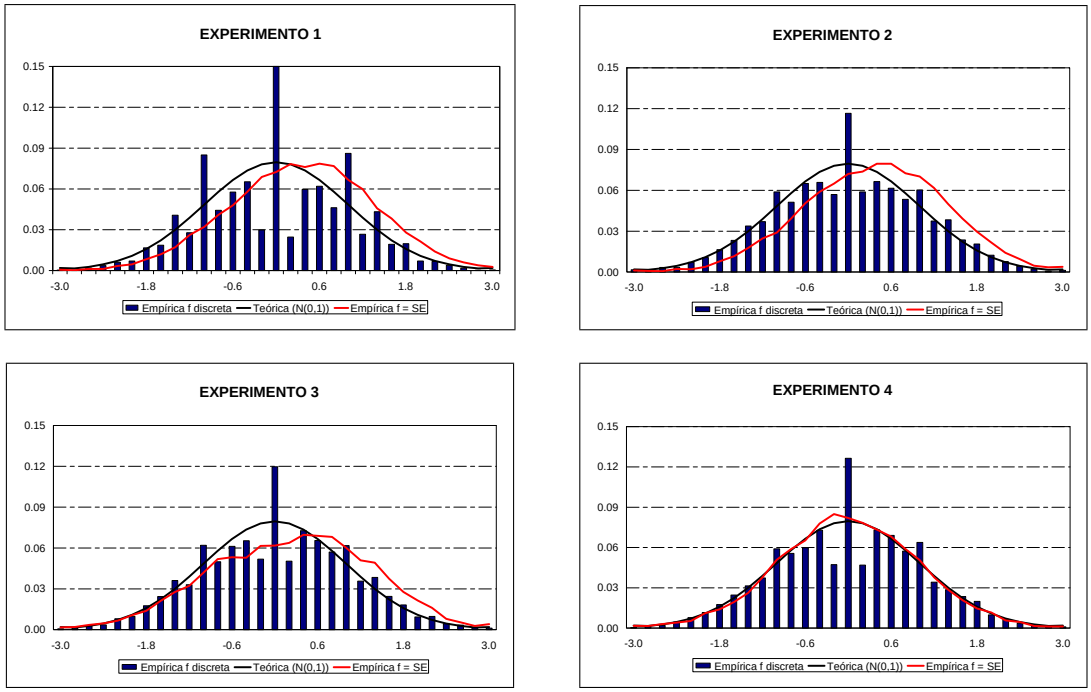


FIGURA 2. DENSIDAD EMPÍRICA ASINTÓTICA ESTADÍSTICO DM BAJO IP
FUNCIONES PÉRDIDA: SE Y DISCRETA. MODELOS ANIDADOS. CASO $\pi = 0,2$.

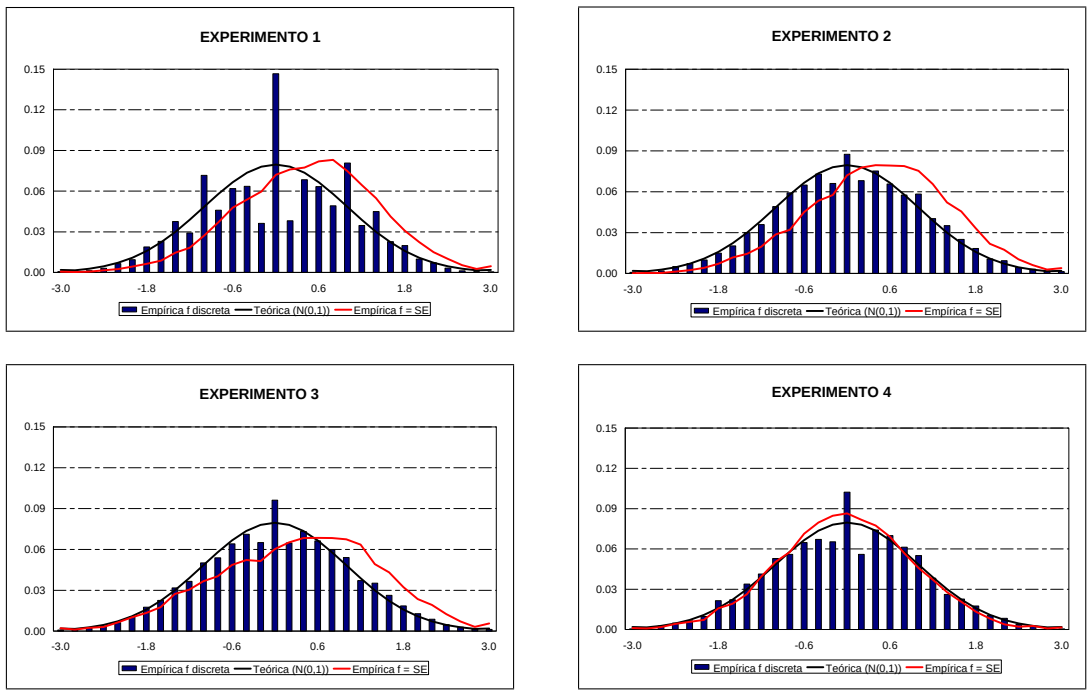


FIGURA 3. DENSIDAD EMPÍRICA ASINTÓTICA ESTADÍSTICO DM BAJO IP
FUNCIONES PÉRDIDA: SE Y DISCRETA. MODELOS ANIDADOS. CASO $\pi = 0,4$.

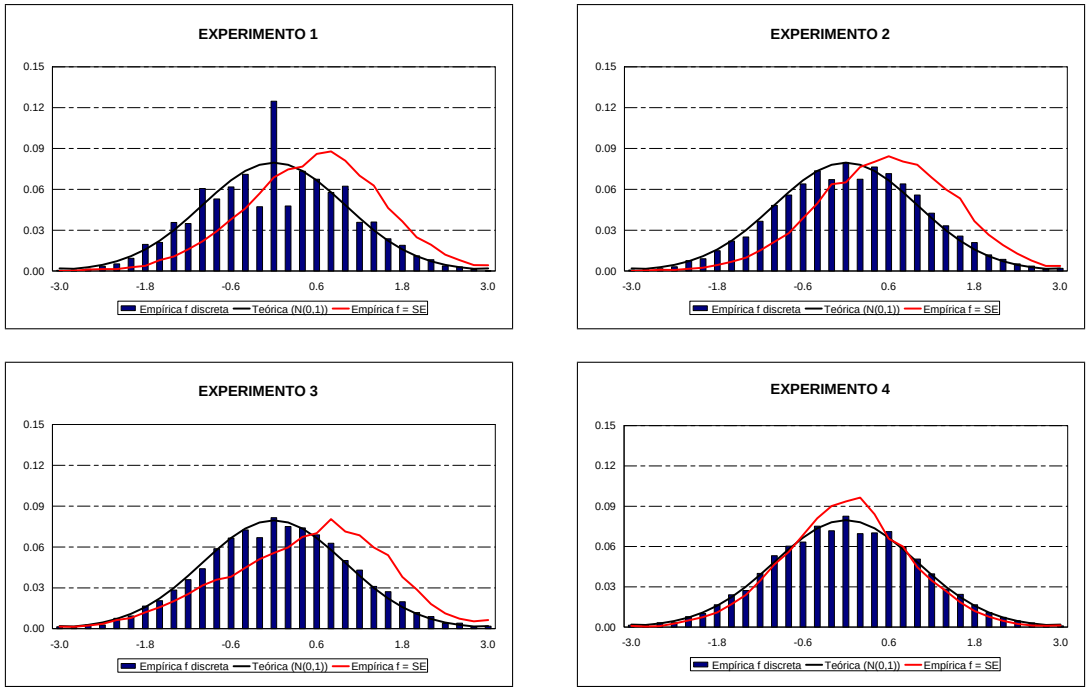


FIGURA 4. DENSIDAD EMPÍRICA ASINTÓTICA ESTADÍSTICO DM BAJO IP
FUNCIONES PÉRDIDA: SE Y DISCRETA. MODELOS ANIDADOS. CASO $\pi = 1$.

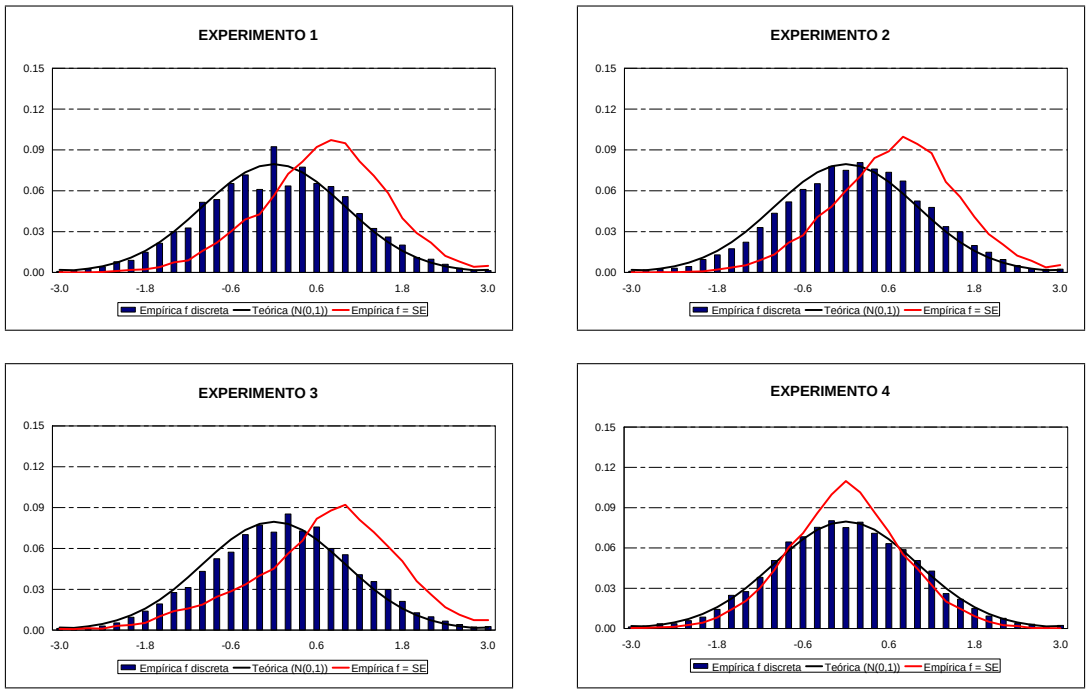


FIGURA 5. DENSIDAD EMPÍRICA ASINTÓTICA ESTADÍSTICO DM BAJO IP
FUNCIONES PÉRDIDA: SE Y DISCRETA. MODELOS ANIDADOS. CASO $\pi = 2$.

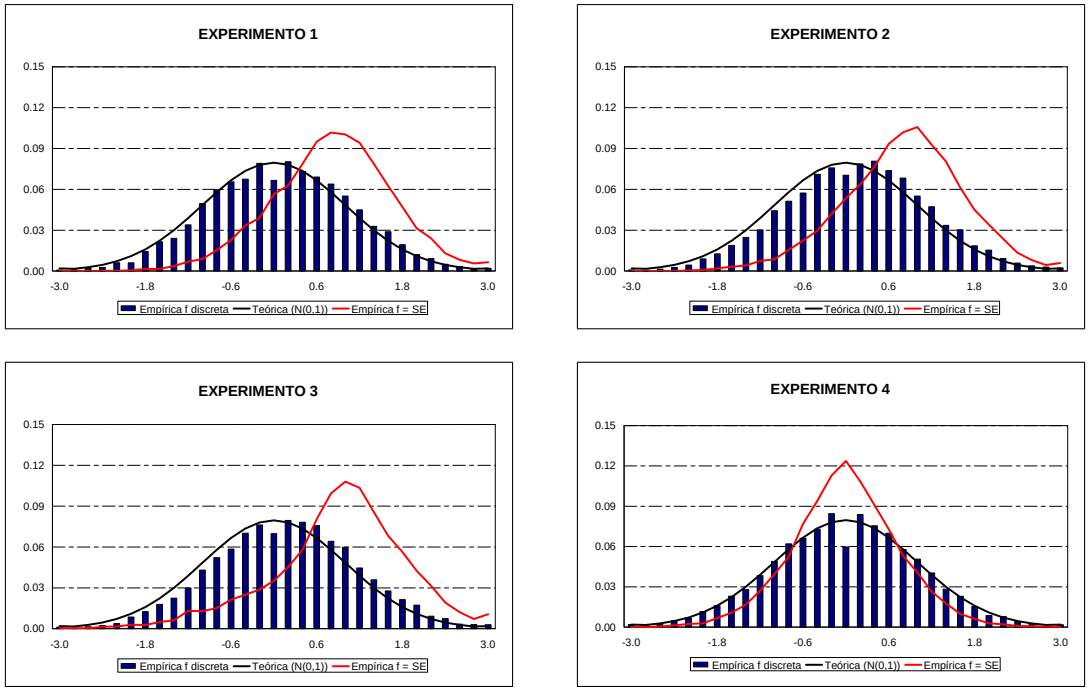


FIGURA 6. DENSIDAD EMPÍRICA ASINTÓTICA ESTADÍSTICO DM BAJO IP
FUNCIONES PÉRDIDA: SE Y DISCRETA. MODELOS ANIDADOS. CASO $\pi = 3$.

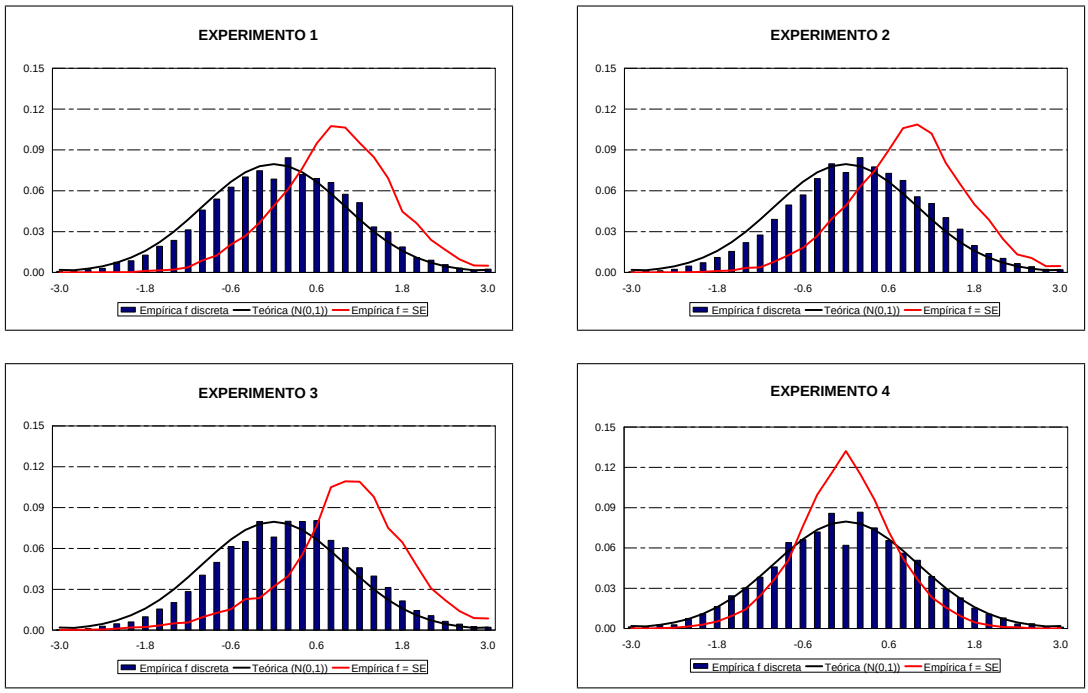
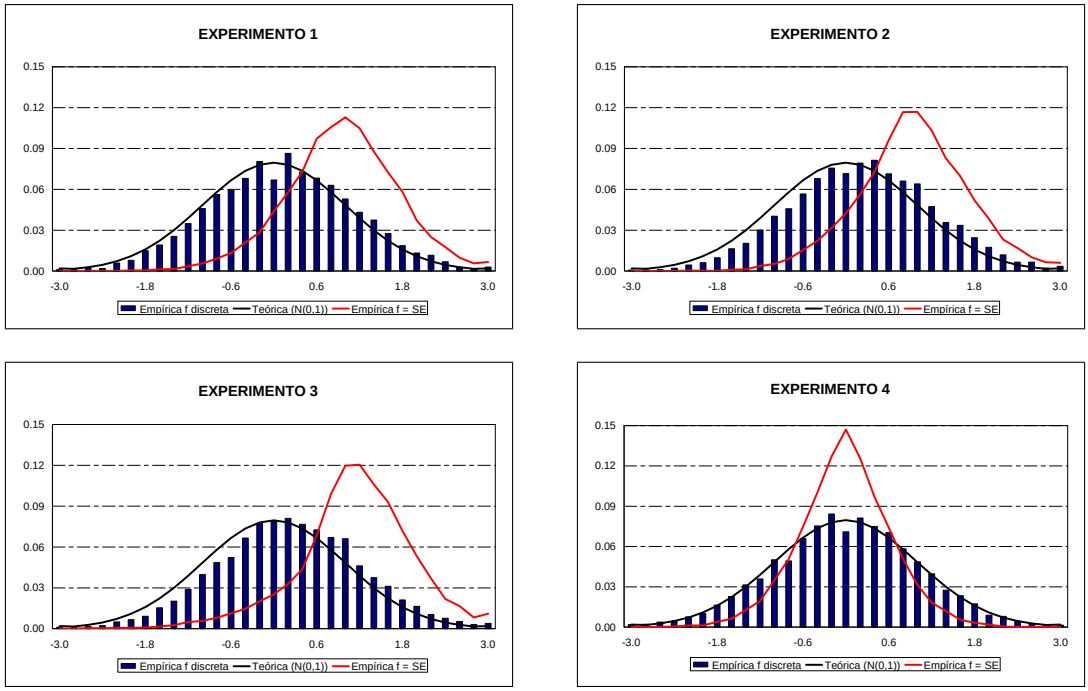


FIGURA 7. DENSIDAD EMPÍRICA ASINTÓTICA ESTADÍSTICO DM BAJO IP FUNCIONES PÉRDIDA: SE Y DISCRETA. MODELOS ANIDADOS. CASO $\pi = 5$.



5. Conclusiones

Hemos tratado de investigar el efecto que tiene la incertidumbre paramétrica (IP) sobre cierto tipo de tests de evaluación/comparación de la capacidad predictiva de un/varios modelo/s. Se trata de tests basados en la media muestral de las pérdidas asociadas a las previsiones de los modelos. Nos hemos centrado en modelos lineales y con variables todas estacionarias. Existe una amplia y reciente literatura cubriendo este contexto, pero dirigida casi siempre a funciones de pérdida continuas y diferenciables, y, normalmente, sencillas (como el cuadrado del error de previsión (SE)). Nuestro interés, sin embargo, radicaba en estudiar el asunto para un tipo particular de funciones "continuas a trozos" − que hemos venido llamando "discretas" −, que habíamos propuesto y motivado en los capítulos anteriores de esta tesis, y que consideramos que pueden tener mucho sentido y utilidad en bastantes aplicaciones. El análisis se ha realizado dividiendo en dos categorías de modelos, no anidados y anidados, ya que tienen implicaciones estadísticas distintas.

Respecto al caso de *modelos no anidados* (uno o varios), se han logrado los siguientes resultados:

a) Hemos demostrado analíticamente (Corolario 1) que, bajo condiciones razonables en la práctica, los tests en cuestión son asintóticamente invariantes (su distribución de contraste no cambia) a la existencia de IP procedente de la estimación de los modelos, siempre que se use una función de pérdida discreta. Estas condiciones son, esencialmente dos: independencia entre los regresores del modelo y sus errores de previsión, y simetría en la distribución de los últimos. Un resultado similar se había mostrado en McCracken (2000) para el valor absluto del error de previsión (AE), aunque exigiendo solamente incorrelación donde nosotros exigimos independencia. Por su parte, el SE verifica la misma propiedad (irrelevancia asintótica de la incertidumbre paramétrica (IAIP)) sin necesidad de establecer supuestos sobre la simetría de la distribución de los errores, y con una condición más débil respecto a la relación entre errores y variables explicativas. Sin embargo, no son habituales este tipo de resultados en la literatura para otras funciones de pérdida, por lo que nuestro hallazgo tiene un interés notable. Además de las hipótesis anteriores, la proposición que hemos obtenido requiere también que la f discreta concreta que se utilice verifique ciertos supuestos. Hemos presentado algunas simulaciones violando tanto estos supuestos como la hipótesis de simetría en la distribución de los errores, y sus resultados sugieren que, aunque el Corolario 1 no es robusto a su incumplimiento, el sesgo asintótico en tamaño de los tests a menudo será muy pequeño. Queda pendiente por identificar qué tipo de definiciones de f discreta y qué situaciones predictivas garantizan dicha conclusión.

La demostración de irrelevancia de la IP bajo f discreta (en las condiciones citadas) se restringe al caso asintótico. Pero los experimentos de Monte Carlo realizados parecen confirmar que la propiedad se conserva también en muestras finitas.

b) Cuando el supuesto de incorrelación entre regresores y errores de previsión no se verifica, la varianza de los tests que consideramos está inevitablemente afectada por la incertidumbre derivada de la estimación de los modelos, sea cual sea la función de pérdida empleada. En el caso del SE, el ejercicio de simulación presentado en West (1996) para el test DM muestra que el tamaño del contraste empeora dramáticamente en caso de no corregir la varianza del test para añadir dicha incertidumbre. Nosotros hemos tratado de cuantificar este empeoramiento tanto en muestras finitas como asintóticamente cuando f es discreta, realizando el mismo experimento de West (1996) pero bajo dicha función, además de bajo el SE y el AE. Los resultados son favorables a la utilización de pérdidas discretas. En caso de usarlas, el sesgo asintótico se mantuvo entre 0 % y 6 % en todos los diseños, mientras con SE y AE lo hizo en niveles que varían entre el 4 % y el 38 %, y entre el 3 % y el 33 %, respectivamente, cuando el tamaño teórico del ejercicio era 5 %. Por su parte, las pruebas en muestras finitas confirman estos resultados, y sugieren que, para la mayor parte de niveles habituales en la práctica de P (número de previsiones) y R (número de observaciones para estimar el modelo correspondiente a la primera previsión), es posible implementar los tests sin corregir la varianza por la IP si f es discreta, con un coste sobre las propiedades del contraste asumible.

La conclusión obtenida es, de nuevo, muy valiosa. La estimación de la varianza correcta bajo IP (Ω) es un proceso complejo y tedioso (especialmente en funciones no diferenciables), como mínimo, e incluso imposible si el usuario del test no coincide con el generador de las previsiones, porque se requiere conocer los detalles de la estimación de los modelos predictivos utilizados. Así que eludir este cálculo sería una gran ventaja, y, de acuerdo a nuestro trabajo, esto es posible bajo nuestra función de pérdida. En cambio, según los resultados obtenidos por nosotros y en West (1996), este proceder sería totalmente inaceptable si se emplea SE, y no cabe sino emplear la complicada varianza asintótica derivada por West (1996).

En el caso de *dos modelos anidados*,[44] nos centramos exclusivamente en el test DM. En este contexto, el estadístico de contraste puede no converger, hacerlo a una distribución degenerada o a una no estándar, incluso aunque no hubiera estimación porque los parámetros de los modelos fueran conocidos. McCracken (2007) logra derivar la distribución asintótica del contraste bajo ciertos supuestos, entre los que está el de diferenciabilidad de la función de pérdida. La distribución resultante no es estándar y McCracken (2007) facilita los percentiles para la implementación del test. Nuestro trabajo ha consistido en tratar de investigar, a través de experimentos de Monte Carlo, cuánto difiere la distribución asintótica de DM respecto de la $N(0,1)$ cuando f es discreta. Los resultados muestran que tal discrepancia es razonablemente pequeña, en general, sobre todo en las colas de la distribución. Por lo tanto, aunque se comparen dos modelos anidados, el test podría llevarse a cabo tomando $N(0,1)$ como distribución de contraste, algo inviable en caso de emplear pérdida SE.

La *conclusión global* de todo el trabajo es clara: el uso de funciones de pérdida discretas garantiza un grado de robustez elevado en las propiedades de los tests de evaluación (tipo C1) o comparación de capacidad predictiva (tipo DM), propiedad que no se certifica si se implementan con la mayoría de funciones de pérdida, ni siquiera con el SE. Si el usuario eligió una pérdida de las que hemos llamado discretas, sería razonable aplicar esta clase de contrastes despreocupándose del efecto de la IP y de la existencia de anidamiento entre los modelos, es decir, puede implementar los tests con la expresión de la varianza asintótica habitual y tomando $N(0,1)$ como distribución de contraste. Como ya se ha mencionado antes, la ventaja de esta propiedad es doble: por un lado, ahorra el tedioso cálculo de la matriz Ω en el caso de modelos no anidados o el incómodo empleo de una distribución no estándar en modelos anidados. Pero, por otro, permite que un evaluador de previsiones procedentes de distintas fuentes externas pueda realizar su trabajo.[45]

No obstante, aunque el resultado del estudio realizado en este documento supone un apoyo importante al uso de pérdidas discretas, no pensamos que deban elegirse éstas solo en base al argumento de la robustez que generan en los contrastes. Nuestra pretensión simplemente es informar de que, en las aplicaciones en las que dicha función sea adecuada por razones conceptuales, las propiedades del test se conservarán razonablemente bien en los contextos que hemos estudiado.

Finalmente, debe destacarse que los resultados obtenidos sugieren que los tests C1-v1 y C1-v2 del Capítulo 1 de la Tesis y los tests DM y Mult2-aprx del Capítulo 2 pueden implementarse exactamente del modo en que se presentaron entonces, sin incorporar en las varianzas de los estadísticos de contraste la incertidumbre provocada por la estimación de los modelos de previsión, tal y como se había anticipado (pero no argumentado) entonces. Como mínimo, esta afirmación será cierta para los contextos predictivos considerados aquí.

Algunas de las *líneas de investigación* con las que se podría continuar este tipo de trabajo surgen de forma natural. En primer lugar, sería importante caracterizar las condiciones de predicción y las características de la partición de la función discreta bajo las que los Supuestos 3 y 4 sobre ésta y la hipótesis de simetría en la distribución de los errores podrían relajarse sin que el resultado de IAIP se viera casi afectado, en el caso de modelos lineales con variables estacionarias independientes del error de previsión. En segundo lugar, interesaría realizar el mismo tipo de análisis sobre el efecto de la IP con pérdidas discretas pero en otros ámbitos predictivos no tratados aquí. Por ejemplo, se podrían considerar modelos no lineales (comprobar si el Corolario 1 puede mantenerse sin el requisito de linealidad) o modelos con variables integradas y cointegradas (extendiendo los trabajos de Rossi (2005) y Corradi, Swanson y Olivetti (2001), respectivamente, al marco de funciones discretas).

[44] Si existe anidamiento, forzosamente nos referimos a $l \geq 2$ modelos, así que se trata de tests de comparación, y no evaluación (caso, en cambio, admisible en modelos no anidados), de capacidad predictiva. Como la práctica totalidad de la literatura sobre este tema, nos restringimos al caso $l = 2$.

[45] Es habitual que un analista quiera contrastar la igualdad de capacidad predictiva entre las previsiones procedentes de distintas instituciones, en cuyo caso solo conoce las previsiones, y no los detalles de los modelos de predicción.

Referencias

[1] Berkowitz, J. y Giorgianni, V. (2001). Long-Horizon Exchange Rate Predictibality?, *The Review of Economics and Statistics* 83, 81-91.

[2] Clark, T. y McCracken, M.W. (2001). Tests of Equal Forecast Accuracy and Encompassing for Nested Models, *Journal of Econometrics* 105, 85-110.

[3] Clark, T. y West, K.D. (2007). Approximately Normal Tests for Equal Predictive Accuracy in Nested Models, *Journal of Econometrics* 138, 291-311.

[4] Corradi, V., Swanson, N. y Olivetti, C. (2001). Predictive Ability with Cointegrated Variables, *Journal of Econometrics* 104, 315-358.

[5] Davidson, R. y Mackinnon, J.G. (2004). Econometric Theory and Methods. Oxford University Press (New York).

[6] Diebold, F.X. y Mariano, R. (1995). Comparing Predictive Accuracy, *Journal of Business and Economic Statistics*, 13, 253-263.

[7] Ericsson, N.R. (1992). Parameter Constancy, Mean Square Forecast Errors, and Measuring Forecast Performance: an Exposition, Extensions and Illustration, *Journal of Policy Modeling*, 14, 465-495.

[8] Harvey, D.I., Leybourne, S.J. y Newbold, P. (1998). Tests for Forecast Encompassing, *Journal of Business and Economic Statistics*, 16, 254-259.

[9] McCracken, M.W. (2000). Robust Out of Sample Inference, *Journal of Econometrics*, 99, 195-223.

[10] McCracken, M.W. (2004). Parameter Estimation and Tests of Equal Forecast Accuracy between Non-Nested Models, *International Journal of Forecasting*, 20, 503-514.

[11] McCracken, M.W. (2007). Asymptotics for Out of Sample Tests of Granger Causality, *Journal of Econometrics*, 140, 717-752.

[12] Pesaran, M.H. y Timmermann, A. (1992). A Simple Nonparametric Test of Predictive Performance, *Journal of Business and Economic Statistics* 10, 461-465.

[13] Pesaran, M.H. y Timmermann, A. (1994). A Generalisation of the Non-parametric Henriksson-Merton Test of Market Timing, *Economics Letters* 44, 1-7.

[14] Rossi, B. (2005). Testing Long-Horizon Predictive Ability with High Persistence, and the Meese-Rogoff Puzzle, *International Economic Review*, 46, 61-92.

[15] West, K.D. (1996). Asymptotic Inference about Predictive Ability, *Econometrica*, Vol. 64, Issue 5, 1067-1084.

[16] West, K.D., McCracken, M.W. (1998). Regression Based Tests of Predictive Ability, *International Economic Review*, 39, 817-840.

CONCLUSIONES

Esta Tesis Doctoral se ha enmarcado dentro de la literatura sobre contrastes para evaluar o comparar la capacidad predictiva de conjuntos de previsiones. La peculiaridad es que trabajamos sobre este tipo de tests introduciendo una clase de funciones de pérdida discontinua en sus argumentos, que denominamos "discreta". La motivación de dicha función y sus bondades conceptuales se presentaron, por ejemplo, en la Introducción de la Tesis (con más detalle, en el Capítulo 1). El desarrollo de este trabajo ha mostrado que, además, la implementación de los contrastes bajo este tipo de funciones genera beneficios técnicos de gran relevancia, que se resumirán posteriormente.

Las **preguntas a resolver** o cuestiones objeto de estudio en la Tesis eran las siguientes:

a) Respecto a los contrastes para la evaluación de capacidad predictiva de un único conjunto de previsiones (**Capítulo 1**):

Los tests estándar de la literatura relativos a este particular pueden comportarse de forma poco satisfactoria, a nuestro juicio. Las razones de que sus prestaciones no sean mejores son fundamentalmente dos. Por un lado, que la pregunta implícita sobre la que deben pronunciarse es ambigua ("¿son las previsiones útiles/valiosas?"). Por otro, que los procedimientos de contraste desarrollados no utilizan una función de pérdida, sorprendentemente. El primer obstáculo es inevitable, pero el segundo abre un campo de acción para la mejora. Entre los tests habituales de la literatura se pueden destacar el Test Binomial (B), el contraste de Henriksson-Merton (1981) (H-M), el de Pearson (Tabla de Contingencia (TC)) o el test de Pesaran-Timermmann (1992) (P-T).

Nuestra pregunta era si podríamos desarrollar contrastes sencillos para evaluar capacidad predictiva, introduciendo una función de pérdida. Por determinadas circunstancias, la función de pérdida natural en este contexto es, precisamente, la discreta. El objetivo es presentar la especificación de algunos tests y analizar su comportamiento, en comparación con los citados anteriormente.

b) Respecto a los contrastes para la comparación de precisión predictiva entre dos conjuntos de previsiones alternativos (**Capítulo 2**):

En este caso, existe un test de referencia, el presentado por Diebold y Mariano (1995) (DM). Es válido asintóticamente, bajo condiciones muy generales, y puede aplicarse prácticamente con cualquier función de pérdida. Sus propiedades en muestras finitas han sido evaluadas con técnicas de Monte Carlo en muchos trabajos, con diseños variados de los experimentos, pero, obviamente, no bajo una función de pérdida discreta. En general, el test DM funciona razonablemente bien en muestras finitas, pero en casos de muestra muy corta y autocorrelación en las pérdidas, presenta un notable sesgo (por exceso) en tamaño.

Las cuestiones que surgen para análisis son dos:

1. ¿Cuáles son las propiedades de tamaño y potencia del test DM en muestras cortas si se implementa con una función de pérdida discreta?.

2. El uso de una función de pérdida discreta garantiza que las pérdidas asociadas a cada conjunto de previsiones siguen una distribución exacta de probabilidad perteneciente a la familia Multinomial.[1] ¿Podemos utilizar esta propiedad para desarrollar algún contraste en este contexto de pérdidas discretas que pudiera competir con DM, y mejorar sus resultados en algún caso concreto?.

c) Respecto a los contrastes de los Capítulos 1 y 2 bajo incertidumbre paramétrica (IP) (**Capítulo 3**):

Los contrastes anteriores se derivaron y evaluaron bajo el supuesto, poco realista en la práctica, de que las previsiones no proceden de un método que requiera estimar parámetros desconocidos. Las distribuciones que se obtuvieron originalmente para los contrastes son, en general, incorrectas cuando aparece IP (en cuanto a sus parámetros caracterizadores −modelos no anidados− o en cuanto a la familia de la distribución −modelos anidados−)). West (1996), McCracken (2000) y McCracken (2007), entre otros, han aportado teoría estadística para la corrección del problema, que resulta compleja en cualquier caso. Pero, obviamente, no existe ningún resultado o análisis a este respecto para el caso particular de que los tests se estén implementando bajo una función de pérdida discreta.

[1] Para ser precisos, son las frecuencias asociadas al conjunto de pérdidas las que siguen una distribución de familia Multinomial.

Las preguntas de interés en este contexto son obvias (y es especialmente relevante responderlas en concreto para el test DM, aplicado bajo pérdida discreta):

1. ¿Cómo adaptar los tests analizados (implementados con función de pérdida discreta) al contexto de IP?.

2. ¿Cuál es el efecto de la IP sobre las propiedades de los contrastes si éstos no se corrigen?. Esta cuestión es trascendente, porque la corrección de los contrastes para incorporar la IP es compleja en todos los casos, y en determinadas circunstancias, imposible. Por tanto, la magnitud de la distorsión que se produce en el tamaño del test si se ignora la IP constituye una información de gran valor.

Las **conclusiones** obtenidas en el trabajo son:

CAPÍTULO 1:

a) Identificamos y mostramos el tipo de errores de comportamiento de los tests tradicionales que pueden considerarse inaceptables. Así, P-T y TC $m \times m$ (versión de TC con una partición genérica formada por $m > 2$ regiones) concluyen que las previsiones son útiles en casos en que la correlación entre datos y previsiones es negativa. Por su parte, los tests H-M, B y TC 2×2 cometen el error de asignar la misma valoración a conjuntos de previsiones con idéntica capacidad de predecir el signo del dato, pero que generan errores de previsión de magnitud bien diferente.

b) Especificamos contrastes sencillos que incorporen la función de pérdida discreta como criterio de valoración de previsiones. Se trata de los tests C1 (con dos versiones, C1-v1 y C1-v2), C2 y C3. Los dos primeros contrastan la posición (identificada por la media poblacional) de la distribución de las pérdidas generadas, utilizando su media muestral como estadístico de contraste. C1 (en sus dos versiones) se deriva de forma sencilla en base a propiedades conocidas sobre convergencia de frecuencias muestrales, y tiene validez asintótica. C2 se especifica en base a la distribución exacta que verifican las pérdidas (familia Multinomial). Por su parte, C3 es un contraste paramétrico especial de Razón de Verosimilitudes (también de validez asintótica), adaptado de una propuesta original de Robertson y Wright (1981).

c) Llevamos a cabo ejercicios de simulación para estudiar el funcionamiento de los tests que proponemos, en relación al de los tests habituales de la literatura. La comparación más relevante es la que se establezca respecto a los tests TC $m \times m$ y P-T, que son los que admiten una partición del dominio de los datos y previsiones flexible (igual que C1, C2 y C3), que es el caso de interés, desde nuestro punto de vista. La ambigüedad de la hipótesis nula que subyace a los contrastes impide un análisis estándar de tamaño y potencia. No obstante, los resultados de los experimentos permiten obtener las siguientes conclusiones:

– Los tests C1, C2 y C3 no cometen ninguno de los errores descritos en a).

– C1 y C2 son más potentes que TC $m \times m$ y P-T en casos en que las previsiones son claramente útiles (correlación entre datos y previsiones positiva y muy elevada). Por su parte, C3 se comporta de manera similar a P-T en estas situaciones.

– En aquellos casos para los que la correlación entre datos y previsiones es intermedia (digamos, que el coeficiente de correlación es positivo pero inferior a 0,7), C1 y C2 deciden de un modo mucho más razonable que P-T y TC $m \times m$, al ser sensibles a la definición de utilidad especificada por el usuario.

d) En base a lo anterior, **aconsejamos** el *uso de los contrastes C1 o C2* para evaluar la utilidad asociada a un conjunto de previsiones. Por razones de coste computacional, es preferible C1, sin duda. Entre sus dos versiones, no se observan discrepancias significativas en sus resultados. El test C1-v2 está obtenido bajo una derivación más correcta que C1-v1, pero éste es aún más sencillo en términos de cálculo.

CAPÍTULO 2:

a) En relación con la primera de las cuestiones planteadas arriba para desarrollar en este capítulo, se concluye que:

– En general, el tamaño en muestras finitas del test DM bajo función discreta reproduce aproximadamente el mismo patrón identificado (por ejemplo, en el artículo original de Diebold y Mariano (1995)) bajo pérdida cuadrática (SE), a saber: el tamaño es bastante exacto, salvo en casos en que las muestras sean muy cortas (no superiores a 16 datos) y, además, éstas presenten autocorrelación en las pérdidas asociadas a las previsiones (en principio, cuando el horizonte predictivo es superior a uno). En esta clase de situaciones, el test es sesgado en tamaño (por exceso). Para una autocorrelación de orden uno, el tamaño de DM es aproximadamente 30% y 20% en casos de $T = 8$ y $T = 16$ (T: longitud muestral), respectivamente, cuando el nivel de significación del contraste era 10%.

– Una discrepancia de cierto interés que identificamos respecto a las propiedades del contraste DM entre su aplicación bajo función continua y su aplicación bajo función discreta es el diferente nivel de robustez del tamaño y potencia del test a la aparición de atípicos o a un grado elevado de no normalidad de los errores de previsión. Siempre que se haya implementado con función discreta, el contraste es robusto a dichas cuestiones (las razones son obvias), mientras su funcionamiento se distorsiona drásticamente por ellas si la pérdida definida era continua. Ésta es una *primera ventaja técnica* del uso de este tipo de criterios de valoración de previsiones, si bien su relevancia no es, ni mucho menos, de la magnitud de la otra ventaja técnica que presentaremos después.

b) En relación a la segunda cuestión, presentamos un contraste basado en el conocimiento de la familia de la distribución asociada a las pérdidas discretas generadas (Multinomial), cuyo enfoque, en cualquier caso, es muy similar al de DM (en cuanto a hipótesis nula y estadístico de contraste). La diferencia es que el test que proponemos (Mult2) utiliza una aproximación a la verdadera distribución del estadístico (que no es conocida con exactitud porque los parámetros de la distribución Multinomial de las pérdidas son desconocidos) que debe ser más cercana a ella que la distribución asintótica de DM. También presentamos la versión asintótica de Mult2 (Mult2-aprx), más similar aún a DM que el propio Mult2, pero conservando alguna diferencia relevante respecto a aquel.

Realizamos simulación de Monte Carlo para comprobar si se obtiene algún beneficio por emplear Mult2 o Mult2-aprx, en vez de la versión discreta de DM, evidenciando las conclusiones a continuación:

– Mult2 obtiene los mejores resultados. La potencia de éste es muy similar a la de DM, pero Mult2 mantiene un tamaño más exacto en todas las simulaciones realizadas, bajo distintos escenarios de autocorrelación y correlación cruzada entre las pérdidas y bajo varias definiciones alternativas de la función de pérdida discreta. Las diferencias son especialmente importantes en los casos de autocorrelación en las pérdidas y muestras pequeñas. Por ejemplo, para el tipo de situación descrita arriba, en la que el tamaño de DM alcanzaba el 30% ($T = 8$) ó 20% ($T = 16$), Mult2 presenta tamaños en torno a 13% (y también Mult2-aprx, pero éste ofrece, en general, peores registros de potencia que DM).

– Además, Mult2 puede computarse siempre, mientras DM podría plantear en algunas ocasiones problemas de negatividad de la varianza del estadístico de contraste, en casos de autocorrelación en las pérdidas, si bien esto sucedió en nuestras simulaciones con una probabilidad siempre inferior al 6% en longitud muestral $T = 8$, y con una probablidad insignificante en longitudes superiores.

– La contrapartida de Mult2 es su alto coste computacional cuando el número de valores posibles de la función de pérdidas (J) es alto o el tamaño muestral grande.

– Nuestra **recomendación** para aquellas aplicaciones de predicción donde proceda incorporar la función discreta es *utilizar Mult2 siempre que la muestra contenga menos de 20 pérdidas, y éstas presenten autocorrelación.* En el resto de casos, podría emplearse el habitual test DM.

c) Finalmente, detectamos ciertos casos peculiares, en los que el tamaño del test de comparación predictiva implementado bajo función de pérdida discreta (Mult2, Mult2-aprx, DM) puede presentar sesgo (por defecto) elevado en muestras muy pequeñas. Caracterizamos dichas situaciones en el trabajo. Se trata de contextos predictivos en los que la probabilidad de que sean exactamente iguales las pérdidas asignadas a las dos previsiones alternativas es muy alta (superior a 0,75 en $T = 8$, y a 0,90 en $T = 16$) y la muestra contiene menos de 20 datos. En cuanto la longitud muestral supera los 20 datos, el problema se anula. Asímismo, el sesgo tiende a desaparecer cuanto más fina es la partición y mayor es J.

CAPÍTULO 3:

Se obtienen las siguientes conclusiones sobre el efecto de la IP en los contrastes de evaluación y comparación de precisión predictiva, cuando la función de pérdida es discreta. Nos limitamos al marco de los modelos de previsión lineales y cuyas variables son estacionarias. De entre los considerados en este trabajo, los tests para los que aplican los resultados que se enumeran a continuación son B, P-T, C1-v1, C1-v2, C2, DM, Mult2 y Mult2-aprx, que son los que contrastan la esperanza matemática de las pérdidas y lo hacen a través de su media muestral, caso en el que se centra la literatura estadística sobre la incorporación de IP en tests de habilidad predictiva.

a) En modelos de previsión **no anidados**:

1. En primer lugar, se obtiene la expresión analítica general de la matriz que constituye el gradiente de la función de pérdida respecto al vector paramétrico a estimar (matriz F, en la notación de McCracken (2000), que es la referencia de interés en nuestro caso). Una vez se tiene dicha expresión, basta seguir el

teorema de McCracken (2000), previa estimación de las matrices oportunas con los datos muestrales, para corregir por la IP los contrastes implementados bajo función de pérdida discreta (tanto los correspondientes al Capítulo 1 como los correspondientes al Capítulo 2). La expresión en cuestión puede encontrarse en el resultado (13) del Capítulo 3, en su versión más general, y en el Lema 1 de dicho capítulo, en una versión que aplica bajo el cumplimiento de ciertos supuestos adicionales sobre la función de pérdida discreta particular elegida, supuestos poco restrictivos en la práctica. El problema que emerge es que la expresión obtenida es compleja, y requiere el conocimiento de la distribución de probabilidad del error de previsión y de los regresores, o bien hacer supuestos a ese respecto.

2. En segundo lugar, se demuestra analíticamente que, bajo cierto conjunto de condiciones y en el contexto predictivo descrito (modelos lineales, con variables estacionarias, no anidados), la IP no genera ningún efecto sobre la distribución asintótica de los contrastes a los que nos venimos refiriendo (denotamos esta situación por las siglas IAIP (irrelevancia asintótica de la incertidumbre paramétrica)), si éstos se implementan bajo una función discreta que verifique cierta restricción sobre su estructura. El resultado completo se presenta en el Corolario 1 del Capítulo 3 y constituye *una de las aportaciones más relevantes de la Tesis*. La IAIP se alcanza también, dentro del mismo tipo de modelos de previsión y bajo un conjunto de condiciones similares al que nosotros imponemos, cuando la función de pérdida es el error de previsión al cuadrado (SE) o el valor absoluto de dicho error (AE), pero el resultado no es extensible a muchas otras funciones. Además, en nuestro caso, nos estamos refiriendo a toda una clase de funciones de pérdida.

La demostración de irrelevancia de la IP bajo función discreta se restringe al caso asintótico. Hemos realizado experimentos de Monte Carlo en muestras finitas, y parecen confirmar que la propiedad se conserva intacta para longitudes muestrales pequeñas.

3. Entre las condiciones exigidas para el cumplimiento de IAIP en los contrastes de evaluación o comparación predictiva, tanto bajo función discreta como bajo SE y AE, destaca la que exige que no exista correlación contemporánea entre errores de previsión y variables explicativas del modelo. A través del mismo diseño de simulación propuesto por West (1996), hemos estimado el efecto de la IP sobre el test DM cuando no se verifica dicha condición, aplicando dicho contraste tanto con una función de pérdida discreta, como con SE y AE. Los resultados son favorables a la utilización de pérdidas discretas. En caso de usarlas, el sesgo asintótico se mantuvo entre 0% y 6% en todos los diseños, mientras con SE y AE lo hizo en niveles que varían entre el 4% y el 38%, y entre el 3% y el 33%, respectivamente, cuando el ejercicio fijaba un tamaño teórico de 5%. Por su parte, las pruebas en muestras finitas confirman estos resultados, y sugieren que, para la mayor parte de niveles habituales en la práctica de P (número de previsiones) y R (número de observaciones para la estimación del modelo correspondiente a la primera previsión), es posible implementar los tests sin corregir la varianza por la incertidumbre paramétrica siempre que la pérdida elegida sea discreta, con un coste razonablemente pequeño sobre las propiedades del contraste.

b) En modelos de previsión **anidados**:[2]

Este caso era especialmente complejo de tratar, porque a diferencia de lo que ocurre en los no anidados, no existe teoría estadística sobre IP para este tipo de modelos si la función de pérdida aplicada en el contraste es no diferenciable. De este modo, nuestra esperanza fundamental era que las pruebas de simulación mostraran que el efecto de la IP sobre las propiedades de los contrastes de comparación de capacidad predictiva aplicados bajo función de pérdida discreta fuera pequeño. Efectivamente, los resultados de los ejercicios de Monte Carlo realizados sobre el test DM y Mult2-aprx sugieren dicha conclusión. Lo contrario ocurre si se utiliza la pérdida SE (en este caso, debe aplicarse la teoría desarrollada por McCracken (2007) para el test DM en situaciones de IP y modelos anidados, según la cual el test DM sigue una distribución asintótica no estándar, que, además, depende del grado de IP). Por ejemplo, contabilizando los tres tipos de experimentos que hemos llevado a cabo y todos sus diseños alternativos, el tamaño en muestras finitas del test DM sin corrección alguna por la IP se mantiene prácticamente en todos los casos entre 5% y 9% si se implementó con pérdida discreta, cuando el teórico era 10%, mientras el tamaño asintótico medio se sitúa en 8,6% (frente al 3,1% que se obtiene si se implementa DM bajo SE, sin corregir por la IP).

[2] Obviamente, este contexto exige que existan dos modelos de previsión, por lo que las conclusiones a continuación aplican a tests del Capítulo 2.

Resumiendo, la **conclusión general del Capítulo 3** es que los tests sobre habilidad predictiva que contrastan la esperanza matemática de las pérdidas a través de su media muestral presentan un *grado de robustez elevado a la IP, siempre que se implementen bajo función de pérdida discreta*. Esta afirmación se restringe al marco de los modelos lineales cuyas variables son estacionarias, y, aplica, por ejemplo, al test P-T, al test DM y, de entre los contrastes que habíamos propuesto nosotros en los Capítulos 1 y 2, a todos los que habían resultado especialmente exitosos en el contexto de no IP, es decir, las dos versiones de C1, Mult2 y Mult2-aprx. Con esto, quedan contestadas las cuestiones 1 y 2 que se plantearon al principio de este apartado, respecto al Capítulo 3. El beneficio derivado del hecho de que la IP distorsione escasamente las propiedades teóricas de los contrastes con función discreta es doble:

i) En el caso de modelos no anidados: se evita tener que realizar el cálculo de la verdadera varianza asintótica del contraste. Dicho cálculo es tedioso siempre, pero a veces incluso imposible de llevar a cabo, si el usuario del test no es conocedor de todos los detalles de la estimación. Incluso si falla el supuesto de incorrelación contemporánea entre errores de previsión y regresores, el sesgo que se produce en el test implementado con pérdida discreta por ignorar la IP sería asumible; en cambio, es inaceptable con pérdidas continuas estándar, como SE o AE.

ii) En el caso de modelos anidados: en principio, no se necesita conocer la verdadera distribución asintótica del contraste, probablemente no estándar. Desarrollar dicha teoría parece un reto de elevada complejidad. En cambio, con otras funciones de pérdida, no queda otro remedio que esperar a que aparezca la teoría estadística adecuada, o, aplicarla si ya existe, pero generando, probablemente, un coste sobre la comodidad de uso del test, al tratarse de distribuciones no estándar, que, además, varían según el grado de IP (entre otras cosas). Proceder en esos casos usando la distribución original del contraste tiende a generar distorsiones en las propiedades de los tests de magnitudes drásticas, a diferencia de lo que ocurre con una pérdida discreta.

Todas estas conclusiones relativas al Capítulo 3 conforman la notable *ventaja técnica* de la función de pérdida discreta, a la que nos habíamos referido al principio de este apartado, y constituyen, probablemente, la *aportación fundamental de la Tesis*.

El trabajo realizado en esta Tesis deja abiertas algunas **líneas de investigación para el futuro**:

Esto se cumple, especialmente, en lo que concierne al Capítulo 3. En primer lugar, sería relevante tratar de extender sus resultados (los analíticos y las inferencias obtenidas a través de ejercicios de simulación) al contexto de modelos no lineales y de modelos con variables integradas y cointegradas. Es posible que adaptar el Corolario 1 para el caso de variables no estacionarias sea poco factible, pero creemos que lo contrario puede decirse respecto a la no linealidad. Además, interesaría especialmente comprobar *qué grado de robustez a la finura de la partición* con la que se define la función discreta presentan aquellas conclusiones del capítulo que se obtuvieron por simulación. Por otro lado, el resultado analítico del Corolario 1 utiliza un supuesto sobre la función de pérdida discreta elegida, que añade cierta estructura en ella. Hemos realizado algunas pruebas (que se muestran en la Tesis) sobre la robustez de la propiedad IAIP al incumplimiento de dicho supuesto, y son favorables, pero dichas pruebas se han limitado a casos muy sencillos y acotados. Sería interesante caracterizar el conjunto de condiciones de la situación predictiva que permitirían incluso relajar el supuesto.

Además de las posibles extensiones del Capítulo 3, creemos que sería razonable trabajar sobre los contrastes de evaluación o comparación de habilidad predictiva introduciendo una *función de pérdida que fuera una mezcla entre la discreta y otras continuas*. Estamos pensando en funciones que sigan discretizando el dominio del par dato-previsión en cuadrantes y limiten el rango de las pérdidas numéricas asociadas a cada uno de los cuadrantes (como en nuestras funciones discretas), pero que, dentro de ellos, se especifiquen funciones continuas del error de previsión. Conceptualmente, estas funciones serían aún de mayor interés en muchas aplicaciones que las puramente discretas, porque siguen permitiendo que el usuario asigne las pérdidas y lo haga con flexibilidad, que se incluya el signo de la previsión dentro del criterio de valoración y se puedan introducir asimetrías, pero, además, serían más precisas en la evaluación de las previsiones del mismo cuadrante (lo que redundaría en un aumento de la potencia de los contrastes). El trabajo a desarrollar se encuadraría, probablemente, en una línea de investigación diferente a la planteada en esta Tesis, porque ya no se podría aplicar el enfoque bajo el que se propusieron contrastes alternativos a los de la literatura (el soporte de las pérdidas dejaría de ser finito), y, además, es muy posible que se perdieran las propiedades bondadosas que tiene la función discreta sobre los contrastes en contextos de IP.